Journal of Nusantara Studies 2025, Vol 10(2) 342-367
Universiti Sultan Zainal Abidin
ISSN 0127-9386 (Online)
http://dx.doi.org/10.24200/jonus.vol10iss2ppx342-367

SENTIMENT AND EMOTION IN MALAY NEWS: A COMPREHENSIVE ANALYSIS USING SENTIMENT ANALYSIS

Mohd Aftar Abu Bakar, *Wan Nurul Huda W Mamat Saufi & Noratiqah Mohd Ariff

Department of Mathematical Sciences, Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
*Corresponding author: nurhuda193@gmail.com

Received: 01.03.2025 Accepted: 07.06.2025

ABSTRACT

Background and Purpose: Emotional framing of news can shape public perception and behaviour. This study examines sentiment in Malay-language headlines from Berita Harian (April 2021–April 2023) to reveal underlying emotions, recurring themes, and societal implications.

Methodology: This research collected headlines from the Berita Harian archive, then applied tokenization, stop-word removal, and normalization. A pre-trained Malay sentiment transformer assigned initial labels (positive, negative, neutral), and a manually verified subset was used to train a Support Vector Machine (SVM). Model performance was measured on a test set via accuracy, precision, recall, and F1-score. Word clouds and count plots highlighted frequent sentiment features.

Findings: The SVM achieved high precision and recall for positive sentiment (0.87/0.85) but lower recall for neutral (0.62), indicating challenges in neutral detection. Dominant topics included COVID-19, PRU15, and *mangsa*.

Contributions: By applying transformer labeling with SVM classification, this work extends sentiment analysis to Malay news media. It informs journalists and policymakers about emotional framing in Malaysian headlines.

Keywords: Sentiment analysis, news headlines, TF-IDF features, Support Vector Machine (SVM), Malay language.

Cite as: Abu Bakar, M. A., W Mamat Saufi, W. N. H., & Mohd Ariff, N. (2025). Sentiment and emotion in Malay news: A comprehensive analysis using sentiment analysis. *Journal of Nusantara Studies, 10*(2), 342-367. http://dx.doi.org/10.24200/jonus.vol10iss2pp342-367

1.0 INTRODUCTION

In today's modern era, news media is deemed to have significant power to impact public opinion and social standards. This is believed to be true by the way news articles, especially their headlines, can greatly influence how readers view and emotionally react to the information presented. It is important to grasp the sentiment behind these headlines, as it offers valuable insights into the emotional aspect of news reporting and its potential impact on the public. Sentiment analysis, a specialized field of natural language processing (NLP), provides effective tools for automatically identifying and examining the emotions expressed in written text. With thorough examination of news headlines, researchers can develop a more profound comprehension of public sentiment and the underlying emotions that are presented in news reporting.

Due to the widespread use of the English language in global conversations, there has been a strong focus on sentiment analysis studies conducted on texts written in English (Chintalapudi et al., 2021). However, this emphasis overlooks the importance of other languages and cultures. This study focuses on analyzing the sentiment of news headlines written in Bahasa Melayu, specifically from Berita Harian, an established Malaysian newspaper. The decision to use Bahasa Melayu is relevant due to its official status in Malaysia and widespread usage in media, education, and daily communication. It is evident that there is a significant lack of computational studies that specifically analysed sentiment in Bahasa Melayu, despite its widespread usage. This research aims to address this knowledge gap by studying the sentiment patterns found in Malaysian media. Specifically, it focuses on analysing headlines from Berita Harian to gain a deeper understanding of how news content in Bahasa Melayu impacts its readers.

Despite the increasing relevance of Malay media, studies applying sentiment analysis to Bahasa Melayu news content remain limited, particularly in the context of formal headlines. Most existing research focuses on English-language texts or informal Malay content such as tweets, leaving a gap in understanding how sentiment is conveyed in traditional Malay news reporting.

This study addresses the research gap by offering three key contributions:

(1) It applies a pre-trained Malay sentiment transformer in combination with TF-IDF and Support Vector Machine (SVM) to classify news headlines into positive, negative, or neutral sentiments.
(2) It establishes a structured preprocessing and sentiment-labelling pipeline specifically tailored for the Malay language, addressing challenges such as tokenization, stop word removal, and informal language structures.
(3) It presents empirical insights into the sentiment distribution of Berita Harian headlines, highlighting recurring themes such as *'covid19'*, *'pru15'*, and *'mangsa'*, and their broader emotional and societal implications.

By focusing on structured formal headlines in Bahasa Melayu—an area largely overlooked in computational linguistics—this study expands sentiment analysis research into under-resourced languages. The integration of a transformer-based Malay sentiment model with TF-IDF and SVM provides a novel methodological framework that combines deep contextual understanding with classical machine learning. Supported by a manually labelled dataset, this hybrid approach enhances the reliability of sentiment classification in formal news content and contributes meaningfully to multilingual NLP research.

2.0 LITERATURE REVIEW

2.1 Text Mining in Malay Language

Text mining involves several types of advanced methods and techniques aimed at extracting valuable insights from textual data. In the context of the Malay language, text mining has been applied in various ways, including extracting compound nouns and normalizing text from social media. Abdul Karim Mohamad et al. (2020), the authors employed text mining techniques to perform sentiment analysis on Malay-language tweets. By applying decision tree classifiers and utilizing a manually labeled dataset, they demonstrated how social media content can be processed and classified based on sentiment categories (positive, negative, neutral). This work highlights the effectiveness of text mining in extracting meaningful patterns from unstructured Twitter data and addresses the challenges of working with low-resource languages like Malay.

New developments in Malay language processing have also led to improvements in sentiment analysis, a crucial application of text mining. Bakar et al. (2019) demonstrated the development of sentiment analysis tools specifically designed for Malay text, pointing out the possibilities of applying NLP techniques to cater to local needs. A recent study by Ying et al.

(2020) delved into the application of advanced machine learning techniques in sentiment analysis for Malay language. The research highlighted the utilization of deep learning models to process Malay-language data, demonstrating the integration of advanced techniques in this field. These studies highlight the increasing relevance of text mining in the Malay language and the notable progress in creating NLP tools specific to this linguistic context.

2.2 Sentiment Analysis for News Headlines

Looking at the emotions and attitudes expressed in media content through sentiment analysis of news headlines can provide valuable insights. News headlines play a crucial role in summarizing articles and influencing how readers think about the content. Researchers have utilized advanced methods, including transformer language models that have been fine-tuned to detect sentiment and label emotions, and subsequently to classify news headlines according to sentiments and fundamental emotions such as anger, disgust, fear, joy, sadness, and surprise (Rozado et al., 2022, Zheng, 2023). These models have proven to be highly successful in capturing the complex emotional content found in headlines, allowing for a more detailed comprehension of the sentiments expressed.

Sentiment analysis is a flexible instrument that can be applied to news headlines in a wide range of fields, such as finance, cybersecurity, and research on social media. As an example, researchers have utilized sentiment analysis to derive useful insights from financial news headlines, make predictions about stock market trends, and assess public sentiment during major events such as the COVID-19 pandemic (Aslam et al., 2020). Within the context of cybersecurity, sentiment analysis plays a crucial role in monitoring the public's response to security breaches and cyber threats. By doing so, it provides valuable insights into the level of trust that individuals place in digital platforms. In addition, when conducting research on social media, it is possible to gain insights into how online communities react to breaking news and how these reactions change over time by analyzing the sentiment of news headlines.

An interesting study conducted by Wongso et al. (2017) looked into the classification of news articles written in the Indonesian language, which exhibits linguistic similarities with Malay. The study utilized a Multi Naive Bayes classifier on a dataset obtained from a local Indonesian news site, resulting in an impressive precision and recall score of 98.4%. This study shows the power of machine learning techniques in categorizing and examining news content in languages other than English, highlighting the importance of sentiment analysis in non-English situations. In a related development, Hossain et al. (2021) conducted sentiment analysis on Bengali newspaper headlines using SVM, Logistic Regression, and Boosted Tree classifiers.

Their findings showed promising accuracy in classifying positive and negative sentiments, even in the face of limited NLP resources for Bengali. These studies illustrate the effectiveness of machine learning techniques in analyzing news content across different low-resource languages. They also emphasize the growing scholarly interest in sentiment analysis beyond English, reinforcing the relevance of applying such techniques to Malay-language news content in this study.

2.3 Analyzing Sentiment in the Malay Language

Understanding the emotions expressed in the Malay language can be quite challenging because of its distinct linguistic characteristics. Malay is a language that uses affixes to modify the meaning of words and sentences. The way words are structured linguistically can make sentiment analysis more complex, as the affixes attached to a word can alter its meaning. In addition, Malay texts, particularly those found on social media, frequently incorporate casual language, mix with their languages like English, and make references to cultural aspects that may not be readily understandable to non-native speakers. The complex structures of language present difficulties for conventional NLP tools and models, which may find it challenging to fully understand the complicated structure of Malay text.

Furthermore, the task is made even more complex by the limited availability of resources for sentiment analysis in Malay. Although there are abundant sentiment lexicons and labelled datasets available for English, the availability of such resources for Malay is limited. Due to limited resources, it is necessary to create personalized lists of words that convey sentiment and manually label datasets (Mahadzir et al., 2022). However, this process can be quite time-consuming and require a lot of effort. To solve these issues, this study suggests a methodical strategy to gathering data, preparing it for analysis, and creating a model that is especially suited to the special characteristics of the Malay language. To overcome these limitations, this study employed a Malay sentiment transformer sourced from the Malaya library, an open-source NLP toolkit for Bahasa Malaysia, based on the BERT architecture and specifically trained for sentiment classification tasks.

2.4 TF-IDF and SVM in Sentiment Analysis of News Headlines

TF-IDF (Term Frequency-Inverse Document Frequency) is a crucial technique in the field of natural language processing and text mining. It is used to assess the importance of terms in a document. This approach considers not only the occurrence of a term inside a text but also its infrequency throughout a collection of papers, resulting in a well-balanced assessment of word

significance (Afif, 2024). TF-IDF is a versatile tool that may be used in various text analysis tasks, such as processing news headlines, analyzing sentiment, extracting keywords, and detecting false news (Iqbal et al., 2023). TF-IDF has proven to be an essential tool in news analysis, since headlines are often succinct and carry significant information.

The strength of TF-IDF lies in its ability to handle the complexity and diversity of text data. Researchers have explored various enhancements and adaptations of TF-IDF to cater to different types of text corpora, thereby expanding its applicability beyond traditional text analysis. For instance, Alammary (2021) discusses how modifications to TF-IDF can improve its performance in specific domains, such as news headline processing, where the challenge lies in capturing the subtle connotations of words within a limited word count.

Multiple studies have repeatedly shown that TF-IDF is beneficial in enhancing the performance of different machine learning models. An exemplary instance involves the utilization of SVM for sentiment analysis of news headlines. SVM models can accurately identify the sentiment, either positive or negative, present in news headlines by utilizing TF-IDF properties. The combination of TF-IDF and SVM has proven to be highly effective in sentiment analysis of media, especially in cases where the precise language of headlines significantly influences public opinion (Iqbal et al., 2023).

Moreover, TF-IDF has been effectively integrated into different machine learning models for various tasks beyond sentiment analysis. In news classification, for example, algorithms like SVM and Multi-Layer Perceptron (MLP) have shown remarkable success when applied over TF-IDF features. Mukhtar et al. (2021) highlight how these models, when combined with TF-IDF, can accurately categorize news content, distinguishing between different types of news articles such as politics, sports, and entertainment. The ability of TF-IDF to capture the essence of words within headlines makes it an essential component in the toolkit for news classification tasks.

SVM specifically, have gained recognition for their ability to effectively handle high-dimensional data, which is frequently encountered when using TF-IDF features. SVM's linearity enables it to establish a distinct separation between classes, such as positive and negative attitudes, by maximizing the gap between them. This attribute is vital in sentiment analysis, as the objective is to precisely differentiate between various emotional tones in headlines.

Recent studies by Osmani et al. (2020, 2022) introduced improved models for sentiment analysis, such as DSLDA, ADSLDA, and enhanced Joint Sentiment-Topic (JST) models. These models add extra information like author, date, and helpfulness to improve results and

are mostly used for long English reviews, such as product feedback. While they offer better performance in finding topics and sentiment, they use unsupervised methods and are not designed for short texts or different languages. In our study, we use a supervised method with TF-IDF, SVM, and a Malay sentiment transformer to analyze short news headlines in the Malay language. Our approach focuses on a low-resource language and a formal news setting, which has not been widely explored in past research.

3.0 METHODOLOGY

The study was initiated by a collection of news headlines from Berita Harian, a prominent Malaysian news source. The headlines were organized into a data frame for further research. To improve the quality of the text data, a preprocessing stage was performed after data extraction. As part of this preparation, the text was normalized and punctuation and stop words were removed.

The main goal in conducting an exploratory data analysis (EDA) was to determine the sentiment expressed in Malay news headlines. EDA included the creation of word clouds and count plots to visually represent the most commonly occurring words and sentiment-specific phrases.

To facilitate the training of a machine learning model, feature engineering was implemented by employing the TF-IDF to extract significant features from the text. The dataset was subsequently divided into training and testing sets using a 70-30 split. This methodological methodology ensured a rigorous approach to analysing the mood of Malay news headlines, which laid the groundwork for valuable discoveries and future research opportunities. The SVM model was selected due to its robustness in handling high-dimensional data from TF-IDF vectors. It constructs an optimal decision boundary, or hyperplane, to separate sentiment classes based on the extracted features.

3.1 Support Vector Machine Fundamentals

Support Vector Machine (SVM) is a supervised machine learning technique widely used for classification tasks, including sentiment analysis. SVM is known for its ability to classify data with high accuracy, especially when there is a clear linear separation between classes. In this context, SVM is used to classify Berita Harian headlines into positive, negative, and neutral sentiment categories.

The core concept of SVM involves constructing a hyperplane in a high-dimensional space that serves as the decision boundary between classes. The hyperplane acts as a separator

that maximizes the margin between different classes. In this sentiment analysis, these classes correspond to positive, negative, and neutral sentiments.

In an n-dimensional space, SVM represents each data point as a vector, where each feature is a coordinate. The goal of the algorithm is to find a hyperplane that best separates the data into their respective classes. In the context of binary classification, the hyperplane divides the input space into two halves, each representing a class label. The function that defines the hyperplane is represented as:

$$f(x) = w^T x + b \qquad (4)$$

where w is the weight vector (which is normal to hyperplane), x is the input feature vector, and b is the bias term. The function $f(x)$ represents the hyperplane that defines the two regions that assist in the classification of the data set. The hyperplane geometrically divides the space into two sections, each corresponding to a distinct category of data under two class labels. A data point "a" is classified into one of the regions based on the value of $f(a)$. If $f(a) > 0$, it is classified within one region; if $f(a) < 0$, it is classified inside a different region. Assume the input data comprises n data vectors, each represented by $x_i \in R_n$, where i=1, 2, …, n. Designate the class label to be applied to the data vectors for supervised classification as y_i, where '+1' represents one category of data vectors and '-1' signifies the alternative category. The dataset can be geometrically partitioned using a hyperplane. Since the hyperplane is represented by a line it can also be presented by:

$$w^T x_i + w^T x_i + b \geq 1 \qquad (5)$$
$$w^T x_i + w^T x_i + b \leq -1 \qquad (6)$$

Upon training the SVM model, the steps used are as follows:

Step 1: Extract Model Coefficients
Step 2: Calculate the slope of the hyperplane
Step 3: Derive the intercept b' of the hyperplane which can be extracted from SVM model output
Step 4: Generate x-coordinate for visualization
Step 5: Calculate corresponding y-coordinates
Step 6: Visualization of the decision Boundary.

With the *x*-coordinates and *y*-coordinates obtained, the decision boundary along with the data points, can be visualized to illustrate how the SVM separates different classes in the feature space.

3.1.1 SVM Implementation

After extracting features using TF-IDF, a linear SVM was applied to classify the headlines into three sentiment categories: positive, negative, and neutral. The linear kernel was chosen for its effectiveness in handling high-dimensional, sparse feature vectors.

Once the model was trained on the labelled dataset, the weight vector and bias term were extracted to derive the decision boundary. These parameters were used to compute the slope and intercept of the hyperplane for visualization, illustrating how the model separates sentiment classes based on key TF-IDF features, as shown in Section 4.7.

The model's performance was then evaluated using standard metrics such as precision, recall, accuracy, and F1-score. Figures 1 and 2 provide an overview of the analytical pipeline, from data collection and preprocessing to classification and evaluation.

Figure 1: Process diagram

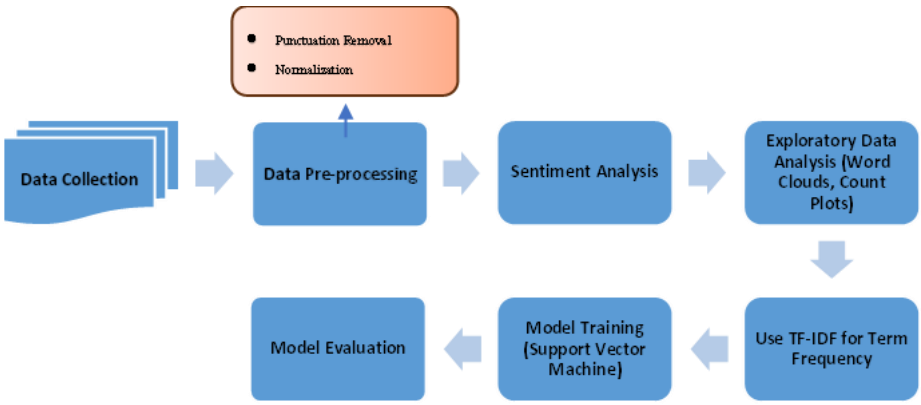


Figure 2: Flow diagram of data pre-processing and model building

4.0 ANALYSIS AND DISCUSSION

4.1 Dataset

This study required the collection of news headlines from Berita Harian between April 2021 and April 2023. The news headlines are taken from the twitter platform, which includes 100,410 tweets, provides insight into the development of the topic of discussion over the past two years. Each tweet was accompanied by a news title, which served as the focused context for the subsequent analysis. The collected tweets were initially aggregated into a list and subsequently converted into a Pandas Data Frame with columns named Datetime and Text. Text designates the textual content of the tweet, whereas Datetime denotes the timestamp at which the tweet is posted.

Name	Description
'Datetime'	The date and time when the tweet were posted
'Text'	The original text content of the tweet
'hash_tag'	The hashtags used in tweet
'Sentiment'	The sentiment label assigned to the tweet
'Text_2'	Processed text content after some preprocessing steps
'Text_3'	Further processed text content (lowercase)
'tokenized'	The tokenized of Text_3 where the text is split into individual words of tokens

Figure 3: Dataset variables

4.2 Data Preprocessing

To preprocess and clean the data, Natural Language Toolkit (NLTK) which is a python library for Natural language processing (NLP) was used for text processing, mainly to remove stop words and stemming. Meanwhile, a regular expression module was used to assist in the removal of alpha numeric characters.

4.3 Removing Punctuation

Punctuation marks, including commas, periods, exclamation points, and question marks, were removed from headlines. Punctuation can introduce unnecessary complexity and noise into text data without providing a significant value for sentiment analysis. For example, punctuation marks might not carry any sentimental value and can disrupt the tokenization process by

splitting meaningful phrases. By removing punctuation, the text becomes cleaner and more uniform, facilitating more accurate tokenization and subsequent analysis. This step helps reduce the dimensionality of the feature space and ensures that the data are more straightforward for the algorithms to process.

4.4 Converting to the Lower-Case Yields

All headlines were converted to lowercase headlines. This step is crucial for maintaining consistency across the dataset because it ensures that words with different cases (e.g., "Kemalangan" and "kemalangan") are treated as the same tokens. Case sensitivity can lead to an inflated vocabulary size in which semantically identical words are treated differently based on their case. Uniformity were ensured by converting all text to lowercase, which helps reduce the complexity of the feature space. This consistency is essential for improving the accuracy and performance of machine learning models as it minimizes the risk of misinterpretation and redundancy in the data.

4.4.1 Remove Stop Words

Stop words are common words that typically do not have a significant meaning in sentiment analysis. Examples of stop words include "dan," "iaitu," "yang," and "di." These words were removed from the headlines to focus on more informative parts of the text. Stop words can add unnecessary noise to the data and often dominate the dataset, thereby overshadowing more meaningful content. By eliminating stop words, we reduced the noise and enhanced the quality of the features extracted from the headlines. This step is essential for improving the efficiency of the model by reducing the number of tokens it must process, allowing it to focus on words that contribute more significantly to sentiment analysis.

4.4.2 Tokenization

Tokenization is the process of splitting headlines into individual words or tokens. This step is fundamental for text analysis as it breaks down the text into manageable components. For instance, the headline "Kemalangan membabitkan empat motosikal dengan sebuah lori berhampiran stesen minyak di Tanjung Musang" would be tokenized into ['kemalangan', 'membabitkan', 'empat', 'motosikal', 'dengan', 'sebuah', 'lori', 'berhampiran', 'stesen', 'minyak', 'di', 'tanjung', 'musang']. Tokenization facilitates the analysis of each word individually and is a crucial step in the feature-extraction process. Proper tokenization ensures that the semantic

structure of the text is preserved, while breaking it down into units that can be effectively processed by machine learning algorithms.

Un named: 0	Date-time	Text	hash_tag	Sentiment	Text_2	Text_3	tokenized
0	2023-04-11 23:44:08 +00:00	#BHKes Kemalangan membabitkan empat motosikal dengan sebuah lori berhampiran stesen minyak di Tanjung Musang.	BHKes	['negative']	Kemalangan membabitkan empat motosikal dengan sebuah lori berhampiran stesen minyak di Tanjung Musang.	kemalangan membabitkan empat motosikal dengan sebuah lori berhampiran stesen minyak di tanjung musang	['kemalangan', 'membabitkan', 'empat', 'motosikal', 'dengan', 'sebuah', 'lori', 'berhampiran', 'stesen', 'minyak', 'di', 'tanjung', 'musang']
1	2023-04-11 23:36:46 +00:00	#BHDunia #Asean Seorang pegawai Imigresen terbunuh, manakala tiga lagi cedera parah.	BHDunia	['positive']	#Asean Seorang pegawai Imigresen terbunuh, manakala tiga lagi cedera parah.	asean seorang pegawai imigresen terbunuh manakala tiga lagi cedera parah	['asean', 'seorang', 'pegawai', 'imigresen', 'terbunuh', 'manakala', 'tiga', 'lagi', 'cedera', 'parah']
2	2023-04-11 23:23:50 +00:00	#BHSukan Bekas juara tiga kali dari Itali itu berkelebihan pada aksi pertama di Estadio da Luz seterusnya memberikan kekalahan pertama buat Benfica dalam kejohanan berkenaan.	BHSukan	['positive']	Bekas juara tiga kali dari Itali itu berkelebihan pada aksi pertama di Estadio da Luz seterusnya memberikan kekalahan pertama buat Benfica dalam kejohanan berkenaan.	bekas juara tiga kali dari itali itu berkelebihan pada aksi pertama di estadio da luz seterusnya memberikan kekalahan pertama buat benfica dalam kejohanan berkenaan	['bekas', 'juara', 'tiga', 'kali', 'dari', 'itali', 'itu', 'berkelebihan', 'pada', 'aksi', 'pertama', 'di', 'estadio', 'da', 'luz', 'seterusnya', 'memberikan', 'kekalahan', 'pertama', 'buat', 'benfica', 'dalam', 'kejohanan', 'berkenaan',]
3	2023-04-11 23:06:58 +00:00	#BHSukan Bayern memerlukan satu daripada kebangkitan terhebat sepanjang zaman lapan hari lagi jika mahu menafikan City laluan ke separuh akhir buat kali ketiga berturut-turut.	BHSukan	['positive']	Bayern memerlukan satu daripada kebangkitan terhebat sepanjang zaman lapan hari lagi jika mahu menafikan City laluan ke separuh akhir	bayern memerlukan satu daripada kebangkitan terhebat sepanjang zaman lapan hari lagi jika mahu menafikan city laluan ke separuh akhir buat kali ketiga berturutturut	['bayern', 'memerlukan', 'satu', 'daripada', 'kebangkitan', 'terhebat', 'sepanjang', 'zaman', 'lapan', 'hari', 'lagi', 'jika', 'mahu', 'menafikan', 'city', 'laluan', 'ke', 'separuh', 'akhir', 'buat', 'kali', 'ketiga', 'berturutturut',]

					buat kali ketiga berturut-turut.		
4	2023-04-11 23:06:20 +00:00	#BHSukan Keputusan suku akhir pertama Liga Juara-Juara.	BHSukan	['positive']	Keputusan suku akhir pertama Liga Juara-Juara.	keputusan suku akhir pertama liga juarajuara	['keputusan', 'suku', 'akhir', 'pertama', 'liga', 'juarajuara',]

Figure 4: Example of pre-processing

4.5 Sentiment Analysis

Sentiment analysis became the main focus following the preprocessing of tweets, of Malay sentiment transformer that capable of categorizing tweets into three different labels – positive, negatives and neutral were use in order to identify sentiments within the tweets written in the Malay language. This transformer made it easier to apply the sentiment label to every tweet, forming the basis for training the machine learning model considered in this study. The sentiment labelling was performed during the data preprocessing and features extraction stages. Each pre-processed headlines were passed through the Malaya sentiment transformer, which automatically assigned a sentiment label. This label was then stored as a new column ('Sentiment') in the dataset. This process was entirely conducted using the Malaya NLP library. By utilizing these sentiment labels, the analysis became more detailed, and predictions made by the machine-learning model are easier to understand. This increased level of detail in the analysis and the improved transparency of the model's predictions allows researchers to gain deeper insights into the underlying emotional dynamics of the conversation.

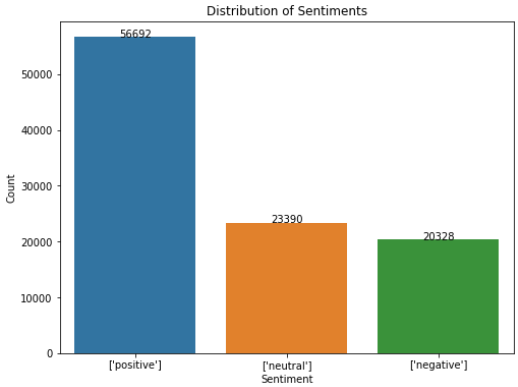


Figure 5: Distribution of sentiment

Figure 5 shows the distribution of sentiments in Malay news headlines, divided into three categories: neutral, positive, and negative. The neutral sentiment dominates with 56,692 occurrences, significantly higher than both positive and negative sentiments. Positive sentiment

headlines amount to 23,390, while negative sentiments are the least frequent, with 20,328 headlines. This imbalance suggests that Malay news headlines may predominantly focus on neutral reporting, with fewer instances of emotional or polarizing language. Such a distribution is important for machine learning models, as the prevalence of neutral data could influence the accuracy and performance of sentiment classification.

4.6 Exploratory Data Analysis

This section contains an in-depth exploration of the dataset to derive significant insights from textual data. Different graphs, such as word clouds and count plots, were created using visualization techniques for positive, negative, and neutral tweets. These illustrations provide an in-depth analysis of the most common terms related to every sentiment category, providing insight into recurring patterns and trends in the dataset. This analysis lays the basis for a deeper assessment of sentiment analysis results by identifying the most frequently occurring words in tweets across various sentiment categories. This helps provide deeper knowledge of the language and popular topics associated with positive, negative, and neutral sentiment results.

4.6.1 Word Cloud

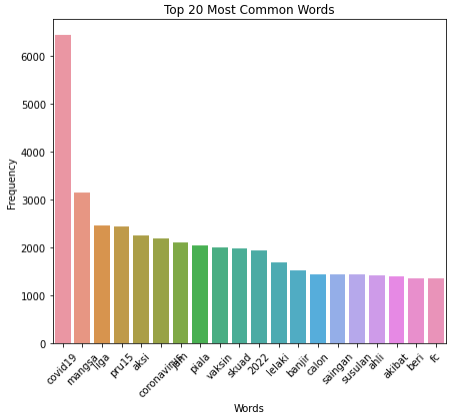
Word clouds (tag clouds) are a type of text data visualization approach that efficiently communicates the main idea of a subject by visually representing words. This method is valuable for analyzing, surveying, and gathering written opinions, among other applications. Word clouds are the most advantageous and uncomplicated means of visualizing text data compared to other techniques. They not only generate a visual representation through language but also communicate additional significance that goes beyond the words themselves. The term 'cloud' is frequently employed to denote the earliest phase of text document analysis, wherein word frequency plays a pivotal role in categorizing or visualizing information.

Figure 6: Word cloud of most common word

Figure 6 displays the most frequently occurring words in Malay news headlines, highlighting prominent terms that reflect major topics and themes. "Covid-19" stands out as the most dominant term, indicating a significant focus on the pandemic in the news. Other related terms such as "coronavirus" and "vaksin" also appear frequently, emphasizing the widespread discussion surrounding the health crisis. In addition to pandemic-related words, terms like "mangsa" , "aksi" , "liga" , and "piala" suggest that news topics also cover various events, including sports and incidents involving victims. Words such as "lelaki", and "skuat" point to discussions involving teams and possibly competitions. The word cloud provides a visual summary of the key subjects covered in the headlines, with the prominence of health and sports-related terms suggesting these areas are of particular interest in Malay news reporting during the dataset's timeframe.

4.6.2 Count Plot

A graph was created to evaluate the distribution of the attitudes in the original dataset. A count plot was generated to illustrate the distribution of sentiments in news headlines. This plot offers a snapshot of how the content of news headlines was expressed over two years. The positive class displayed a significantly greater quantity of news headlines, while the negative and neutral categories consisted of approximately 20,000 headlines each.



Figures 7: Top 20 most common words

Figures 7 shows that the most prominent word in the dataset is "covid19," with a frequency exceeding 6,000 occurrences. This is inline with the findings of the word cloud in Figure 6 where the largest word in covid19. The high frequency of this term reflects the significant impact of the pandemic on various aspects of life and its pervasive presence in the public discourse.

Following ‘covid19’ the word ‘mangsa’ appears with a frequency of around 2,500. This suggests that discussions involving victims, possibly related to the pandemic or other incidents, were also prevalent. Other notable words include ‘polis,’ ‘kes’, ‘coronavirus’, and ‘vaksin,’ each with frequencies ranging from approximately 1,500 to 2,000. These words indicate a focus on law enforcement, specific cases or incidents, and vaccination efforts, all of which were critical topics during the pandemic. The chart also includes words such as ‘mati’, ‘lewat’, ‘calon’, ‘saman’, and ‘sukan’, each with frequencies between 1,000 and 1,500. These words suggest discussions about mortality, delays, political candidates, legal summons, and sporting events. The presence of these words indicates a diverse range of topics, from health and safety to politics and sports.

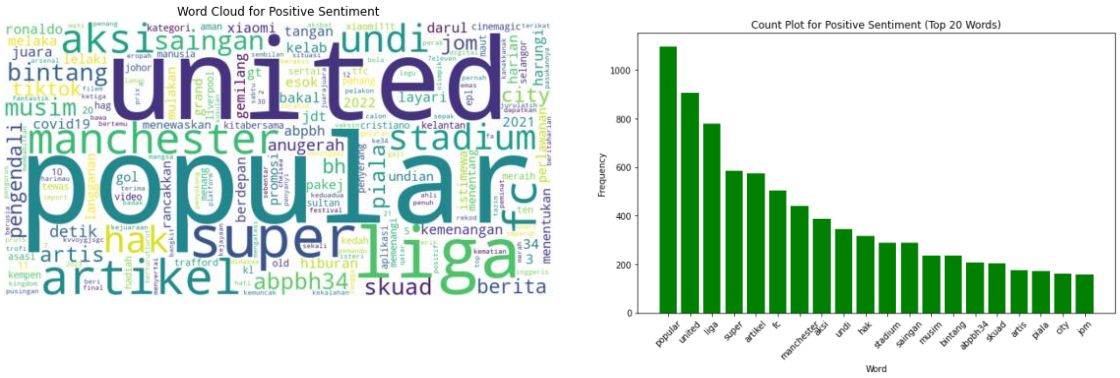


Figure 8: Word cloud and count plot positive sentiment

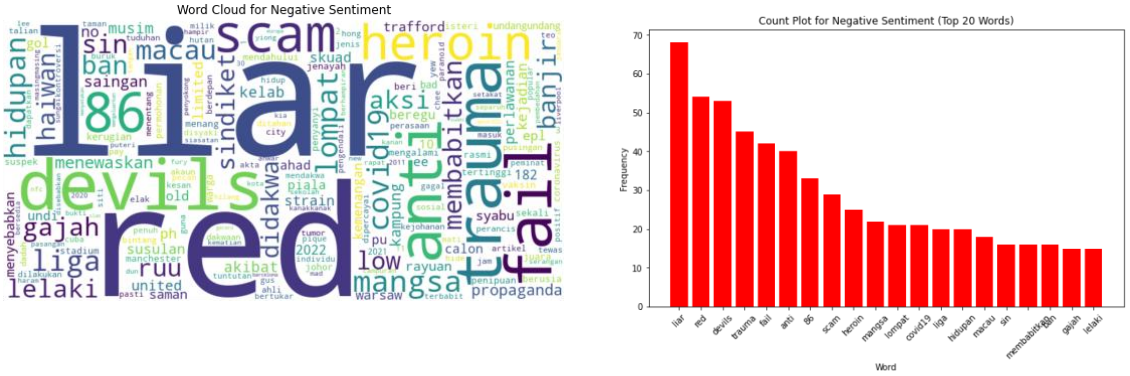


Figure 9: Word cloud and count plot negative sentiment

Figures 8 and 9 show a detailed analysis of both positive and negative sentiments in the dataset, emphasizing different themes via word clouds and bar charts. Figure 8 highlights positive mood through phrases such as ‘popular’, ‘united’, and ‘super’ emphasizing themes of popularity and achievement, especially in sports and globally respected organizations. This indicates that the

dataset comprises positive and motivational information. In contrast, Figure 9 illustrates negative sentiment through notable phrases like 'liar', 'devils', and 'scam' which signify themes of fraud and hardship. The recurrent reference to sensitive subjects such as 'heroin' and 'covid19' signifies involvement with societal issues. The comparison of these data provides a comprehensive perspective on the sentiment landscape, where positivity is associated with achievement and recognition, but negativity is fuelled by dishonesty and controversy. This analysis provides critical insights into public mood, facilitating informed decision-making and alignment with audience ideals.

4.6.3 News Categories

The headlines were categorized into various news categories, and the sentiment distribution within each category was analyzed. Categorizing news headlines into various topics helps us understand the sentiment distribution within each category.

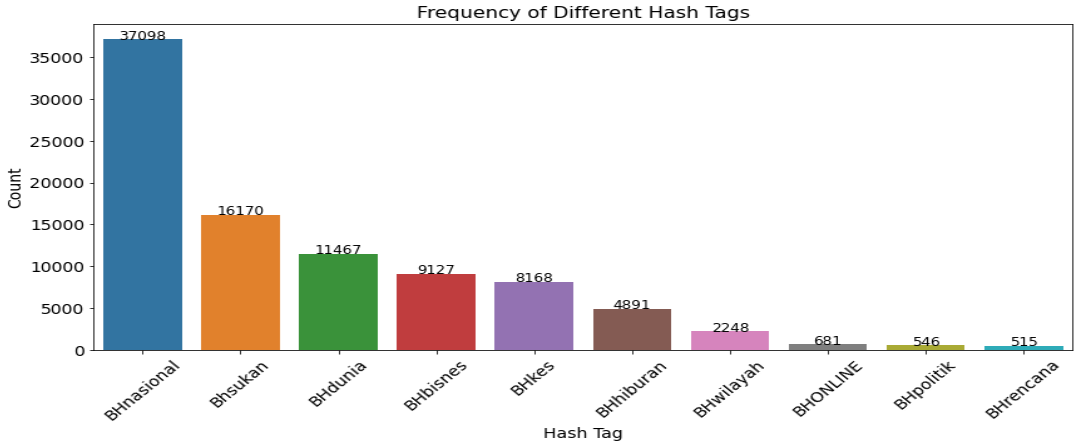


Figure 10: Frequency of news categories

Figure 10 illustrates the distribution of Malay news headlines across several categories, offering valuable insights into the primary areas that influence sentiment analysis. Politics emerges as the dominant category, with 37,098 headlines, making it a crucial focus for sentiment analysis, as political news often provokes strong emotional responses. Sports, comprising 16,170 headlines, represents another significant category where sentiment analysis can capture public reactions, ranging from excitement to disappointment, particularly concerning sporting events or athlete performance. Crime, with 11,467 headlines, is likely to evoke predominantly negative sentiments, making it essential for identifying societal concerns or distress. Entertainment, accounting for 9,127 headlines, provides a varied sentiment

landscape, including both positive and negative emotions, depending on the nature of the news. Economy, at 8,168 headlines, reflects sentiments linked to financial stability, job markets, or economic policy, all of which profoundly impact public sentiment.

The remaining categories Accidents**,** Technology**,** Health**,** Education**,** and Environment, while less represented, still provide important insights into more specialized areas of sentiment, such as responses to technological advancements, health crises, or educational reforms. This distribution highlights key focus areas for conducting sentiment analysis on Malay news headlines.

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

After the preprocessing steps, the text data are ready for feature extraction using TF-IDF. TF-IDF is a technique that is widely used for extracting features from textual data. It assigns weights to words based on their occurrence in a document relative to their frequency across the entire corpus. The TF-IDF score is calculated by multiplying the term frequency (TF) of a word in a document by the inverse document frequency (IDF) of that word across all documents. Term frequency is the number of times a term appears in a specific document, while inverse document frequency is the logarithmically scaled inverse fraction of documents containing the term. This calculation reduces the impact of commonly occurring words across the entire corpus, emphasizing terms that are more specific to individuals.

The TF-IDF score for a term t in a document d within a corpus D is calculated as follows:

Term Frequency:

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ term\ in\ document\ d} \quad (1)$$

Inverse Document Frequency (IDF):

$$IDF(t,D) = log\ log\ \left(\frac{Total\ number\ of\ documents\ in\ corpus\ D}{1+Number\ of\ document\ containing\ term\ t}\right) \quad (2)$$

TF-IDF Score:

$$TF - IDF\ (t,d,D) = TF(t,d) \times IDF\ (t,D) \quad (3)$$

TF-IDF helps highlight important words in a document and is especially useful in tasks such as information retrieval, text mining, and natural language processing. By focusing on unique terms, it improves the differentiation between documents based on their content, making it a valuable tool for various text analysis applications.

In this study TF-IDF score indicates the significance of specific words such as 'covid19', 'mangsa', 'pru15', coronovirus', and 'liga' in the collection of news headline from Berita Harian. The higher TF-IDF score suggests that the word is relevant and unique in the context of the news headlines, with 'covid' being the most prominent based on the score provided. Table 1 shows an analysis of Malay news headlines from Berita Harian with their respective TF-IDF scores for the top 5 words.

Table 1: TF-IDF score

Word	TF-IDF-Score
covid19	0.011
mangsa	0.006
pru15	0.005
coronavirus	0.005
liga	0.004

TF-IDF is employed to extract essential terms in news headlines, aiding sentiment classification and topic categorization. By focusing on terms with high TF-IDF scores, analysts can identify the most relevant words that convey significant information about the content and context of the news. The analysis of Malay news headlines from Berita Harian using TF-IDF reveals significant terms that reflect prevailing themes and topics, contributing to a deeper understanding of public sentiment and interest. The highest TF-IDF score for "covid19" indicates that the pandemic was the most dominant topic during the analyzed period, which aligns with the global health crisis affecting all aspects of society. The frequent mention of "coronavirus" further reinforces the focus on public health concerns, highlighting the ongoing challenges and media coverage related to the pandemic. This widespread attention on COVID-19 evokes a range of sentiments, from fear and concern to hope and resilience, shaping public discourse and emotions. Similarly, the term "mangsa" suggests a focus on incidents involving victims, stemming from coverage of crimes, accidents, or natural disasters, which often elicit negative sentiments such as sympathy or outrage. The appearance of "PRU15" (Malaysia's 15th General Election) reflects significant media attention on political developments, a topic

known to provoke polarized opinions and sentiments. Lastly, the term “liga” highlights the importance of sports, particularly football, which tends to generate enthusiastic responses from the public, from enthusiasm to disappointment. These prevailing themes provide insights into the dominant narratives in Malay news and their impact on societal sentiment.

4.8 Sentiment Analysis using TF-IDF Vectors

Each headline's TF-IDF vector is a sparse representation of term importance, where only nonzero entries are shown. This means that the vector includes only terms that are relevant for that particular headline, not every possible term in the corpus. The TF-IDF vector for a headline includes the TF-IDF scores for multiple terms present in the headline. For example, in the first headline, terms like ‘kemalangan’ and ‘lori’ are part of the TF-IDF vector. The value inside bracket represent the position of the term in the vocabulary, and the TF-IDF scores represent the importance of the terms in the given headline. Table 2 shows a few news headlines with their corresponding TF-IDF vectors. Each headline's TF-IDF vector represents the weighted importance of terms within that specific headline, aiding sentiment classification.

Table 2: TF-IDF vector

News Headline	Sentiment	TF-IDF Vector
"Kemalangan membabitkan empat motosikal dengan sebuah lori berhampiran stesen minyak di Tanjung Musang."	Negative	(0, 81759) 0.36319164412617727, (0, 112804) 0.351681467044376, (0, 35784) 0.10846258007476561, ...
"Seorang pegawai Imigresen terbunuh, manakala tiga lagi cedera parah."	Positive	(0, 91075) 0.37141884702472383, (0, 30745) 0.3193416618612696, (0, 70984) 0.24506989167144996, ...
"Bekas juara tiga kali dari Itali itu berkelebihan pada aksi pertama di Estadio da Luz seterusnya memberikan kekalahan pertama buat Benfica dalam kejohanan berkenaan."	Positive	(0, 22288) 0.3354855666939089, (0, 25670) 0.14525746258717226, (0, 66283) 0.168397133608002, ...
"Bayern memerlukan satu daripada kebangkitan terhebat sepanjang zaman lapan hari lagi jika mahu menafikan City laluan ke separuh akhir buat kali ketiga berturut-turut."	Positive	(0, 105036) 0.36316609449701287, (0, 116618) 0.14903387318068395, (0, 26258) 0.20671344203015776, ...
"Keputusan suku akhir pertama Liga Juara-Juara."	Positive	(0, 33657) 0.596085500819111, (0, 72441) 0.2415449097579942, (0, 110974) 0.3072918915022952, ...

In Table 2, the value in bracket (81759, 112804…) represent the positions of the terms in the vocabulary, and the TF-IDF scores (0.363, 0.352…) indicate the importance of these terms within the headline. The vector reflects that terms like ‘kemalangan’ and ‘lori’ are important

in the context of this headline. This analysis was based on the entire headline text. This represents the importance of certain terms within each headline, considering the entire corpus used for the analysis. This method highlights the key terms in each headline that contribute significantly to the overall text analysis and sentiment detection. By analyzing these vectors, one can identify which terms are most significant in each headline and how they might relate to the overall sentiment and thematic content of the news corpus.

4.9 SVM Visualization

In the context of Malay news headline sentiment analysis, Figure 11 illustrates the classification of data points using a Support Vector Machine (SVM) model. The theoretical basis and implementation process are explained in Sections 3.1 and 2.1.1. The preprocessing steps for this analysis included tokenization, stop word removal, and TF-IDF vectorization, as described earlier in Section 4.5. These steps enabled the extraction of meaningful features that contributed to sentiment classification.

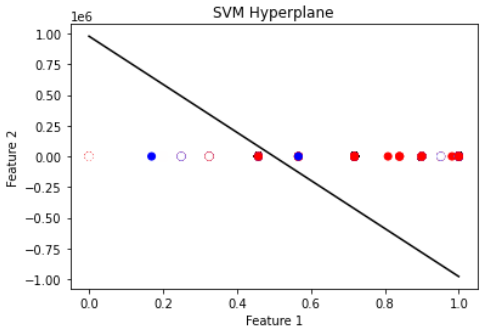


Figure 11: Hyperplane of support vector machine

After feature extraction, SVM constructed an optimal hyperplane, depicted as the black line in the figure that separates the data points into two distinct sentiment categories, as represented by red and blue dots. The points closest to the hyperplane, known as support vectors, are indicated by the solid dots (both red and blue), which play a critical role in defining the margin, the distance between the hyperplane and the nearest data points from each class. Maximizing this margin improves the model's ability to generalize to unseen data. The hollow points, on the other hand, represent correctly classified data points that lie further from the decision boundary and do not influence the hyperplane's position. By constructing the widest possible margin, SVM ensures a clear and accurate separation between the sentiments within the dataset, ultimately improving classification performance.

5.0 MODEL EVALUATION

The assessment of the SVM model's performance is crucial for comprehending its effectiveness in classifying the sentiment expressed in the Malay Berita Harian News headlines. Different performance metrics are used to evaluate the performance of algorithms. Several common performance metrics, including Precision, F1 Score, Accuracy, and Recall, are used to analyze the performance and identify the most effective algorithm for this project. These measures were evaluated using actual and prediction-based metric.

Table 3: Results evaluation metrics

Evaluation Metrics		Actual	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy measure is used to verify that the number of predictions made by a classifier are accurate. However, this measure is not sufficiently convincing when used on dataset that is unbalanced, as it makes inaccurate predictions for the other categories and prioritizes the category with the highest frequency.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (4)$$

where TP is True Positive, TN is True Negative, FN is False Negative, and FP is False Positive. Similarly, Recall, Precision, and F1-Score are used to measure the performance of the classifiers. These are much better than that of accuracy to determine the performance of the algorithms when the dataset is not well balanced, and their formulas are given below:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 - Score = 2 \times (Precision \times Recall)/(Precision + Recall) \quad (7)$$

By examining these metrics, we can gain a more comprehensive understanding of the SVM model's performance on diverse datasets and make informed decisions regarding its application in various contexts.

5.1 Model Performance Metrics

The evaluation of the SVM model's performance was validated through the use of several metrics that offer insight into its classification abilities. These metrics consist of accuracy, precision, recall, and the F-1 score, which are determined for each sentiment category: negative, positive, and neutral. Table 4 shows the classification outcomes.

Table 4: Model SVM accuracy result

Sentiment	Precision	Recall	F1-Score	Sample
Negative	0.78	0.64	0.70	3850
Positive	0.75	0.89	0.82	11085
Neutral	0.70	0.48	0.57	4496
Accuracy	0.747			

Table 4 displays the accuracy result for the SVM model. The model achieved an accuracy of 0.747, indicating that it correctly classified approximately 75% of the news headlines. This suggests that the proposed model is effective for sentiment classification. The precision measures the proportion of correctly predicted positive observations to total predicted positive observations. The model exhibited high precision across all sentiment categories. The recall measures the proportion of correctly identified positive observations to total actual positive observations. The model has high recall for positive sentiment, indicating that it effectively identifies positive headlines; however, recall for neutral sentiment is lower, suggesting potential challenges in accurately identifying neutral sentiment. The F1-score is a harmonic mean that balances precision and recall. Although SVM are widely utilized for sentiment analysis and have demonstrated efficacy in predicting both positive and negative attitudes, the neutral sentiment F1-score suggests potential for enhancement. Their constraints in precisely categorizing neutral attitudes derive mainly from their binary classification framework, the ambiguity of neutral language, and the impact of dataset attributes. In order to capture the complexity of neutral sentiments better, it may be important for researchers to explore alternative models or hybrid approaches as the field of sentiment analysis continues to develop.

5.2 Macro and Weighted Average

In the SVM model performance metric, the macro average indicates how well the model performs across each sentiment category (negative, positive, and neutral), without considering the number of headlines in each category. This is important for understanding the ability of the model to handle each sentiment in a category independently. Meanwhile, the weighted average refers to a performance metric that considers the number of samples in each class, offering a more comprehensive view of overall performance. It weights the contribution of each class's performance by the proportion of samples in that class, meaning that classes with more samples have a greater impact on the final score. The results are shown in Table 5 for both macro and weighted average of the SVM model.

Table 5: Macro and weighted average

Macro Average		
Precision	Recall	F1-Score
0.74	0.67	0.70
Weighted Average		
Precision	Recall	F1-Score
0.74	0.75	0.74

On the other hand, Table 5 shows comparison between the macro and weighted averages for an SVM model's precision, recall, and F1-score in a sentiment classification task. The macro average treats each sentiment class equally, showing how well the model performs across all categories without considering class size, with a balanced F1-score of 0.70. Meanwhile, the weighted average accounts for the number of headlines in each class, giving a more comprehensive view of performance, with a slightly higher F1-score of 0.74, reflecting better overall accuracy due to class imbalance.

6.0 CONCLUSION

This study assessed a TF-IDF + SVM pipeline for classifying 100 k Malay Berita Harian headlines into positive, negative, and neutral sentiment. The model reached an overall accuracy of 74.7 %. Positive headlines were dominated by sports wins and other achievements, negative ones by crime and political scandals, while neutral headlines, the largest share delivered factual updates. Together, these results sketch the prevailing emotional tone of recent Malay news and offer editors and researchers a data-driven view of public mood.

Remaining challenges stem largely from language features, idioms, slang, and code-mixing, that blur sentiment boundaries, especially for neutral content. Future work should expand the corpus and benchmark transformer-based Malay models against the current SVM baseline to see whether deeper architectures can capture subtler cues and raise recall, particularly for neutrality. The present findings nonetheless confirm that the established TF-IDF + SVM pairing is a practical starting point for large-scale sentiment monitoring in under-resourced language.

REFERENCES

Afif, M. (2024). Applying TF-IDF and k-NN for clickbait detection in Indonesian online news headlines. *Journal of Advanced Computing Knowledge and Algorithms, 1*(2), 38-41.

Alammary, A. (2021). Arabic questions classification using modified TF-IDF. *IEEE Access, 9*(1), 95109-95122.

Aslam, F., Awan, T., Syed, J., Kashif, A., & Parveen, M. (2020). Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. *Humanities and Social Sciences Communications, 7*(23), 1-9.

Bakar, N. S. A. A., Rahmat, R. A., & Othman, U. F. (2019). Polarity classification tool for sentiment analysis in Malay language. *IAES International Journal of Artificial Intelligence, 8*(3), 258-263.

Chintalapudi, N., Battineni, G., Di Canio, M., Sagaro, G. G., & Amenta, F. (2021). Text mining with sentiment analysis on seafarers' medical documents. *International Journal of Information Management Data Insights, 1*(1), 100005.

Hossain, M. S., Jui, I. J., & Suzana, A. Z. (2021). *Sentiment analysis for Bengali newspaper headlines* [Unpublished undergraduate thesis]. BRAC University.

Iqbal, B. M., Lhaksmana, K. M., & Setiawan, E. B. (2023). 2024 presidential election sentiment analysis in news media using support vector machine. *Journal of Computer Systems and Informatics, 4*(2), 397-404.

Mahadzir, N. H., Omar, M. F., Nawi, M. N. M., Salameh, A. A., Hussin, K. C., & Sohail, A. (2022). MELex: The construction of Malay-English sentiment lexicon. *Computers, Materials & Continua, 71*(1), 1790–1807.

Mohamad, A. K., Jayakrishnan, M., & Nawi, N. H. (2020). Employ Twitter data to perform sentiment analysis in the Malay language. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(2), 1404–1412.

Mukhtar, R., Iqbal, M., & Faheem, Z. (2021). Pakistani news classification based on headlines. *Pakistan Journal of Engineering and Technology, 4*(4), 79-85.

Osmani, A., Mohasefi, J. B., & Gharehchopogh, F. S. (2020). Enriched latent dirichlet allocation for sentiment analysis. *Expert Systems, 37*(4), e12527.

Osmani, A., Mohasefi, J. B., & Gharehchopogh, F. S. (2022). Weighted joint sentiment-topic model for sentiment analysis compared to ALGA: Adaptive lexicon learning using genetic algorithm. *Computational Intelligence and Neuroscience, 2022*(4), 1-35.

Rozado, D., Hughes, R., & Halberstadt, J. (2022). Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models. *PLOS ONE, 17*(10), e0276367.

Wongso, R., Luwinda, F. A., Trisnajaya, B. C., & Rusli, O. (2017). News article text classification in Indonesian language. *Procedia Computer Science, 116*(1), 137-143.

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, 9*(2), 212-220.

Zheng, X. (2023). Stock price prediction based on CNN-BiLSTM utilizing sentiment analysis and a two-layer attention mechanism. *Advances in Economics Management and Political Sciences, 47*(1), 40-49.