PCMJ

Progress in Computer and Mathematics Journal (PCMJ)
volume 3 [November, 2025]
e-ISSN: 3030-6728
Website: fskmjebat.uitm.edu.my/pcmj

EMOTION CLASSIFICATION BASED ON CRIME NEWS USING SVM MACHINE LEARNING

RABIATULFITRI MOHD ISHAK
Bachelor of Computer Science (Hons.), College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM) Cawangan Melaka, Kampus Jasin, Malaysia
rabiatulfitri00@gmail.com

Article Info

Received: March 2025
Accepted: September 2025
Available Online: November 2025

Abstract

Emotion may be shown in a variety of manners. These include voice, written texts, and facial expressions and movements. Emotions are divided into six separate categories such as fear, joy, sadness, anger, disgust and surprise. Emotion classification in text, particularly in crime-related news articles, is a crucial task for understanding public sentiment. Classification emotion in normal text is complex, and it becomes even more challenging with news text that does not specify the emotions that they convey, adding complexity to the process. Therefore, in order to solve the problem is to develop an emotion classification system for crime news. The dataset is collected from a reliable news source which are The Star, The Sun and NST. The model is developed using Support Vector Machines (SVM), utilizing a dataset scraped from the news website. The system uses Word2Vec for word embedding to capture semantic relationships and contextual meaning in the text. The methodology follows the Modified Waterfall model, which adapts the traditional Waterfall process while allowing for flexibility and iterative feedback during development. This includes phases of requirement analysis, system design, implementation, testing, and deployment. The project follows a linear sequence but allows for feedback loops and adjustments during key phases like testing and evaluation. The system is evaluated using performance metrics such as accuracy, precision, recall, and F1 score. Preliminary results indicate that the model achieves an accuracy of 81%, precision of 76%, recall of 60%, F1-Score of 62% Word2Vec SkipGram compared to Word2vec CBOW. Future work will focus on the user interface with features like detailed visualizations, interactive dashboards and support for multiple languages to greatly enhance the overall user experience and accessibility.

Keywords: Emotion Classification; Crime News; Sentiment Analysis; Text Classification; Support Vector Machine (SVM); Word2Vec; SkipGram; News Analytics; Natural Language Processing (NLP); Modified Waterfall Model; Text Mining; Performance Metrics; Precision; Recall; F1-Score; Machine Learning; Data Scraping; News Sentiment Analysis.

INTRODUCTION

Emotion is an integral part of human communication, expressed through speech, facial expressions, and written text. Understanding emotions in crime-related news articles is particularly significant as it helps gauge public reactions to criminal activities. However,

identifying emotions in news text is complex due to its formal tone and lack of direct emotional markers. The important part of the narrative is the emotion of the stories themselves. The stories typically will evoke all kinds of emotions from various types of stories in various types of people either listeners or readers (Christ et al., 2022). According to Plutchik (2001), emotions are categorized into six different types which are joy, sadness, fear, surprise, anger, and disgust. This statement is to bring out the emotions in the words or phrases that someone is trying to convey through their speech, facial or text.

Recognizing emotions from text is a fundamental yet challenging task for automated systems, crucial for enhancing human-computer interaction. Unlike methods such as analyzing facial expressions or vocal tones, which provide rich contextual cues, a text lacks these non-verbal signals, posing a significant obstacle for emotion recognition systems (Alrasheedy et al., 2022). Crime news often evokes fear, sadness, or anger among readers. This study aims to address this challenge by developing an emotion classification system that analyses crime news and classifies the text into six primary emotions: fear, joy, sadness, anger, disgust, and surprise.

Traditional text-based news, such as that found in newspapers, and online news and media-based news, like that in television, represent two primary forms of news dissemination (Andersen et al., 2016). Research from Burggraaff and Trilling (2017) has shown that online news articles tend to have differences in news values compared to print articles with online news being more likely to be follow-up items and containing variations in references to persons, power elite, negativity, and positivity.

News articles, which typically emphasize reporting events, frequently lack the emotional words and features essential for emotion classification (Kennedy et al., 2012). This absence of emotion-related content in news articles poses challenges in classifying the emotions they evoke. Current methods for text emotion classification may struggle when directly applied to news articles because of this deficiency in emotional content (Li et al., 2016). Studies have shown that emotions play a significant role in news stories, affecting how information is interpreted and perceived by readers (Takatsu et al., 2020).

A narrative consists of a series of actions that are connected both temporally and logically. Although most approaches focus on the role of narrative in knowledge representation it can be argued that narrative is also related to emotions (Akimoto, 2020). Some authors even argue that emotions and narratives have a homologous structure (Pólya & Csertő, 2023).

This project leverages Support Vector Machines (SVM), a robust machine-learning algorithm known for its effectiveness in text classification tasks. Word2Vec SkipGram is used for word embedding to enhance feature representation, allowing the system to capture semantic relationships between words. The system follows the Modified Waterfall Model, ensuring a structured yet flexible approach to development.

LITERATURE REVIEW

Emotion may be shown in a variety of manners. These include voice, written texts, and facial expressions and movements. Emotions are divided into six separate categories. That includes joy, sadness, fear, surprise, anger and disgust (Plutchik, 2001).

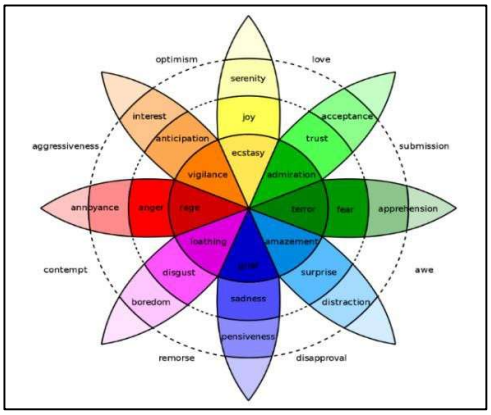

Figure 1 Plutchik's wheel of emotion

This section explained the general explanation of each six basic types of emotion based on Robert Plutchik's wheels. Firstly, **joy** is a pleasant emotion associated with well-being and satisfaction, often expressed through smiling or speaking in an upbeat tone. In contrast, **sadness** can be manifested by crying, being quiet, or withdrawing from others, encompassing feelings such as grief, hopelessness, and disappointment. **Fear,** which can increase heart rate and trigger racing thoughts or the fight-or-flight response, may arise from real or perceived threats. **Disgust**, on the other hand, can be triggered by physical experiences such as seeing or smelling rotten food, blood, or poor hygiene. **Anger** is often shown through facial expressions like frowning, yelling, or even violent behavior. **Surprise,** which can be either pleasant or

unpleasant, might cause someone to open their mouth or gasp and, like fear, can also trigger a fight-or-flight response.

METHODOLOGY

The methodology of this project follows a structured approach to ensure accuracy and efficiency in emotion classification. The Modified Waterfall Model is used, incorporating both structured and flexible elements to allow iterative improvements. This section details the data collection, preprocessing, model training, and evaluation steps.

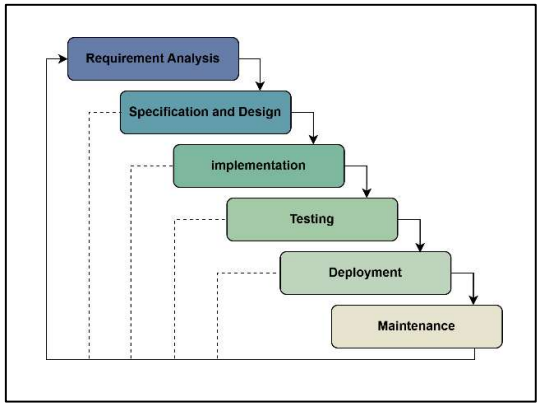

Figure 2 Modified Waterfall Model

Table 1: Methodology Description

Phase	Deliverable	Objective
Requirement Analysis	Problem Statement Objectives Objective Scope and Significance	Objective I
Design	Classification Model Use Case Diagram Flowchart Diagram User Interface	Objective I
Implementation	Cleaned dataset Classification model Web System prototype	Objective II
Testing	Functionality Test Result	Objective III

Requirement Analysis

The requirement analysis starts with determine the problem statement to solve, understanding the project's scope and outlining the objectives that need to be achieve. A total of 146 articles have been downloaded and 76 articles have been cited in this project. Several articles have been reviewed to provide a clear understanding into the project's scope, the problem statement, potential solution and the significance of the project. The collected articles sources are the UiTM online database including IEEE, Springer, ACM, Google Scholar and many others.

Table 2: Requirement Analysis

No.	Objectives
Objective I	To design a system that classifies the emotion in crime news
Objective II	To develop an emotion classification system on crime news using a Support Vector machine algorithm.
Objective III	To evaluate the functionality and the accuracy of the developed system.

Design

In design phase, the focus shifts to creating the blueprint for the system. This phase involves several critical design activities, such as a classification model, use case diagram, flowchart diagram, and user interface, which collectively set the groundwork for the system's development.

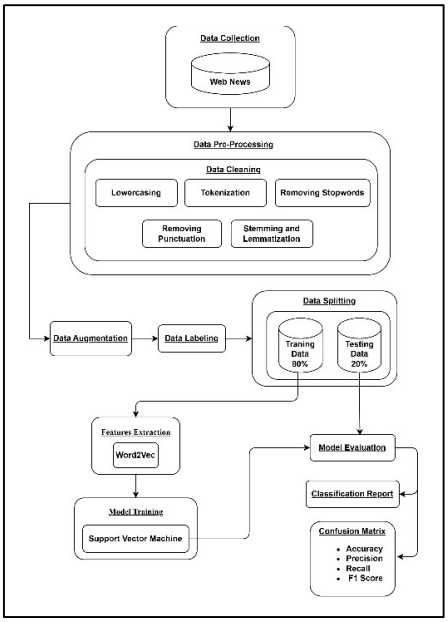

Figure 3 Classification Model

PCMJ
Progress in Computer and Mathematics Journal (PCMJ)
volume 3 [November, 2025]
e-ISSN: 3030-6728
Website: fskmjebat.uitm.edu.my/pcmj

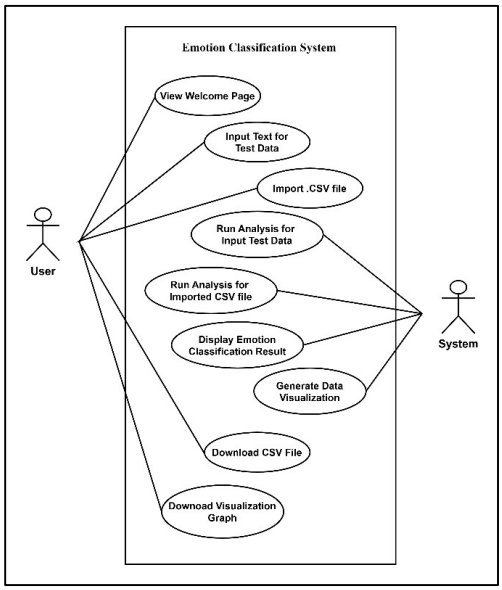

Figure 4 Use Case Diagram

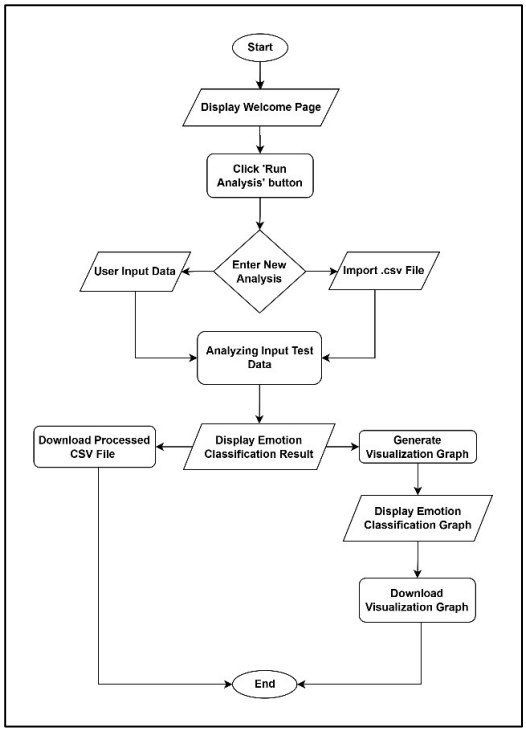

Figure 5 Flowchart Diagram

Implementation

The collected data sources are mainly in news websites. The focusing websites are from The Star, The Sun and NST. A total of 10,000 data are required to collect in two datasets each. Once the data is collected, process to develops a classification model is execute and implement algorithm in this phase. Simultaneously, web development activities are carried out to create the system's front-end and back-end components, which will be the environment in which the model operates.

The emotion classification project involves transforming trained SVM model into a web development. In the beginning, the backend of the application is developed using Python with frameworks which is Flask which handle HTTP requests from the frontend. The pre-trained SVM model is responsible for classifying emotions based on crime news which integrated into the backend. The input from the user which is a crime news article is preprocessed using NLP techniques such as tokenization, stopword removal and lemmatization that facilitated by the NLTK library.

After preprocessing, the text is converted into a vectorized format using Word2Vec which is then fed into the SVM model for emotion classification. The backend then returns the predicted emotion such as fear, sadness, anger, disgust, surprise or joy back to the frontend. The frontend is built using HTML for the page structure where users can input crime news articles in a text area and submit them using a button. CSS is used to style the page and make it visually appealing**.** The final user experience with users receiving instant feedback on the ***e***motion classification of the crime news articles they submit and error handling involved in this phase include Python, Flask, NLTK, HTML, CSS, JavaScript.

Testing

Functionality tests include examining the system for any mistakes or problems with user interaction. The confusion matrix is a table that summarizes the number of true positive, true negative, false positive, and false negative predictions made by the model for each emotion class. A basic structure confusion matrix is made of 2 by 2 as Table 3.

PCMJ
Progress in Computer and Mathematics Journal (PCMJ)
volume 3 [November, 2025]
e-ISSN: 3030-6728
Website: fskmjebat.uitm.edu.my/pcmj

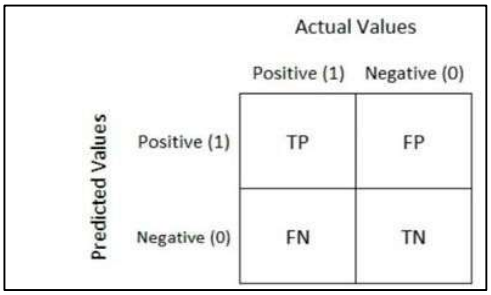

Figure 6 Basic Confusion Matrix

Table 3: Basic Confusion Matrix

Values	Column Header Goes Here
TP	The number of positive samples predicted.
FP	The number of negative samples predicted falsely as positive.
TN	The number of negative samples predicted correctly as negative.
FN	The number of positive samples predicted falsely as negative.

The classification metrics selected for evaluation are F1-score and accuracy. Accuracy represents the ratio of correct predictions to the total number of predictions made. On the other hand, the F1-score, ranging from 0 to 1, incorporates both precision and recall by computing their harmonic mean, providing a balance between the two metrics. (Nasir et al., 2020). The formula for the said metrics is shown in Eq. (1), Eq. (2), Eq. (3) and Eq. (4).

$$\boldsymbol{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad \textbf{1}$$

$$\boldsymbol{Recall} = \frac{TP}{TP + FN} \quad \textbf{2}$$

$$\boldsymbol{Precision} = \frac{TP}{TP + FP} \quad \textbf{3}$$

$$\boldsymbol{F1 - Score} = \frac{2 \times Precision \times Recall}{Precision \times Recall} \quad \textbf{4}$$

DEVELOPMENT

It involves two components, which are back-end and front-end development. These are required in Web Application projects. In back-end development, the section discussed the data collection and data pre-processing. Hence, the SVM algorithm was implemented to construct the model.

Data Collection

Data preparation is a critical step in the data analysis and machine learning pipeline. It involves data collection, cleaning, transforming, feature engineering, data integration, data splitting, data augmentation, validation and organizing raw data to make it suitable for analysis and modeling.

Data Source

The dataset for this study was compiled from three major Malaysian news outlets: The Sun, The Star, and NST (New Straits Times).

Figure 7 The Sun Web News
(*TheSun | Malaysia News: National, World, Viral & Free ePaper*, n.d.)

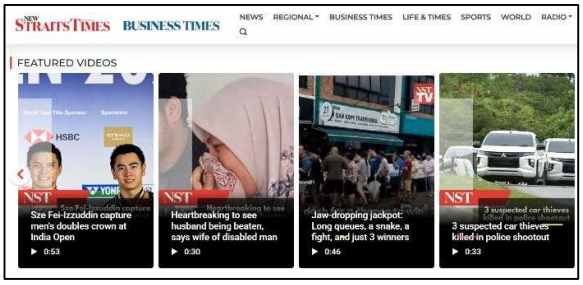

Figure 8 New Straight Time Web News
(*New Straits Times (NST Online) | Malaysia News & World Updates*, n.d.)

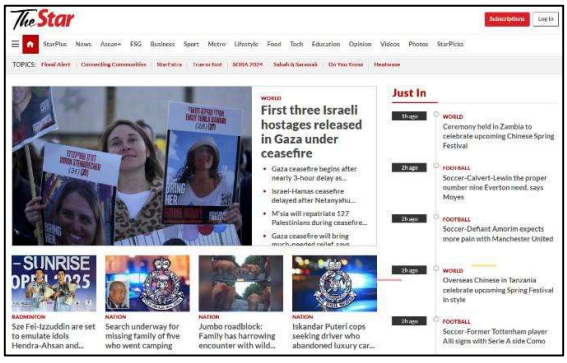

Figure 9 The Star News Website
(*The Star Online | Malaysia, Business, Sports, Lifestyle and Video News*, 2019)

Table 4: Total Data Collection

Sources	Child Abuse	Scams
The Sun	1386	684
The Star	758	467
New Straight Time	217	417
Total Raw Data	**2361**	**1570**

Data Augmentation

Table 5: Total Data

Dataset	Child Abuse	Scams
Raw	2361	1570
Cleaned	2094	1389
Augmented	10470	6945
Cleaned + Augmented	12564	8334
Combined Data	**20898**	

Data Labelling

This project labeled the data using the NRC Emotion Intensity Lexicon by Saif Muhammad (Mohammad, 2020). The lexicon was developed to quantify emotions and sentiment dimensions in textual data. It is widely used in emotion detection and sentiment analysis research. The NRC Emotion Intensity Lexicon is a rich dataset that contains 7477 entries, in which each word is assigned to one of the following eight emotions: fear, anger, sadness, joy, disgust and surprise. These words were assigned intensity values from 0 to 1, with 0 implying that the word has no association with emotion and 1 signifying a strong association. The dataset has three columns: word (the word being analyzed), emotion (the emotion that corresponds to the word), and intensity (the extent to which the word is associated with the emotion).

Table 6: Lexicon Data

No.	Dataset	Child Abuse	Scams
1	outraged	anger	0.964
2	divorce	surprise	0.398
3	terrorism	disgust	0.734
4	criminality	fear	0.642
5	honest	joy	0.303
6	adultery	sadness	0.566

Table 7: Total Word in Lexicon

No.	Emotion Class	Total Data
1	Fear	1763
2	Anger	1481
3	Sadness	1294
4	Joy	1264
5	Disgust	1092
6	Surprise	583

Features Engineering

This involves two main processes: feature selection and feature extraction. This project removed columns such as news_headline and news_url since they do not contribute to emotion classification. The news_content column remains as it provides meaningful textual data for analysis. The dataset was created by merging multiple CSV files into one unified file (*crime_news.csv*), ensuring consistency in processing. TF-IDF (Term Frequency-Inverse Document Frequency) where it converts text into weighted numerical features based on word importance within a corpus.

Next, Word2Vec (SkipGram & CBOW), it generates word embeddings that capture semantic relationships. The SkipGram model predicts context words given a target word, while CBOW predicts the target word from context words.

Classification

Firstly, X_train_skipgram is the subset of the feature set that will be used to train the model. X_test_skipgram is the subset of the feature set that will be used to test or evaluate the model's performance after training. On the other hand, y_train is the target variable for the training set and y_test is the target variable for the test set. Next, test_size=0.2 indicates the value of 20% of the data to the test set and the remaining 80% will be used for training. This is the 80:20 split. Then, random_state=42 is used to control the randomness of the split. Setting it to a specific value like 42 to ensures that the split will be the same every time the code is running.

Table 8: Data Splitting

Total Data	Training	Testing
20898	80%	20%
	16718	4180

Comparisons are made between different feature extraction methods, such as TF-IDF and Word2Vec (SkipGram and CBOW), to determine the most effective approach.

Table 9: Summary Training Result

Embedding	Model	Size of data	Accuracy
TF-TDF	SVM	20898	0.95
SkipGram	SVM	20898	0.83
CBOW	SVM	20898	0.79

Table 10: Testing Result

Embedding	Model	Accuracy	Precision	Recall	F1-Score
TF-TDF	SVM	0.93	0.82	0.77	0.76
SkipGram	SVM	0.81	0.76	0.60	0.62
CBOW	SVM	0.77	0.61	0.54	0.54

RESULT AND DISCUSSION

Functionality Test

This section emphasizes the effectiveness, accuracy and reliability of the implemented features.

Table 11: Sample of Functionality Test

Test Description	User input their particular long text to run the analysis.
Test Input	**Test Data:** a lorry driver was charged at the sessions court here today with two counts of rape and unnatural sex involving a widow with three children inside the vehicle last month. however, the accused, mohd norhisyamudin jamaludin, 44, pleaded not guilty to all charges read to him by the court interpreter before judge osman affendi mohd shalleh. according to the first charge, the father of three was accused of raping the 43-year-old widow inside a lorry parked at taman kota, yong peng, near here, between 7.39pm and 8.30pm on jan 28.
Expected Test Outputs	Anger
Actual Test Results	

Figure 10 Actual Test Output

Table 12: Functionality Test Report

Title	Description
Name of Tester	Salehah binti Hamzah
Date and Time	5/2/2025 4.00-5.00 PM
Location	KPPIM, UiTM Kampus Jasin
Position	Senior Lecturer, UitM Kampus Jasin
Objective	Evaluate the emotion prediction model and evaluate the prediction of system prototype via web platform.
Area Covered	Emotion prediction accuracy, Web deployment completeness
Test Approach	Manual
Total Test Passed	6
Total Test Failed	2

Data Visualization

By leveraging charts, and graphs, this project enhance interpretability, making complex data more accessible and actionable for decision-making. It shows that fear emerges as the most frequent emotion followed closely by sadness. This aligns with expectations for crime news where the events described often evoke fear or sadness in the public. These emotions are typical responses to news of violent crimes, accidents, or tragedies. In contrast, joy and anger are less frequent in the dataset.

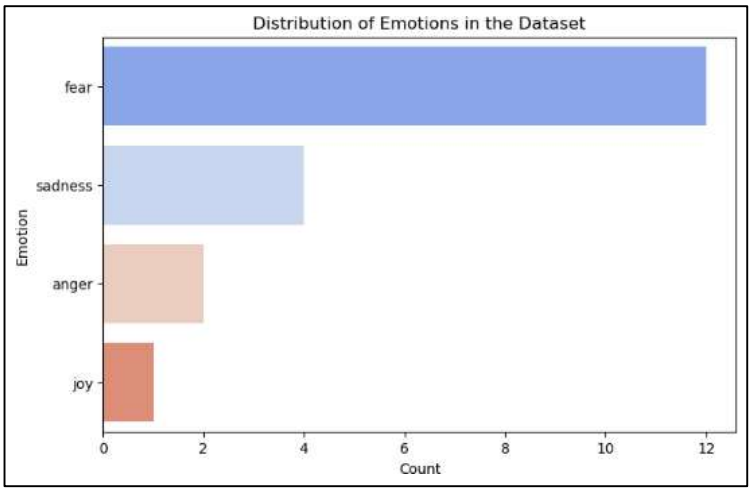

Figure 11 Emotion Distribution

CONCLUSION

This project was designed to analyse crime-related news articles and predict them into specific six basic emotions which are fear, sadness joy, disgust, surprise and anger. The project successfully integrated Natural Language Processing (NLP) techniques to preprocess textual data, extract relevant features and train a classification model. The system was evaluated for its accuracy, and functionality by providing promising results while identifying areas for further enhancement.

Therefore, all project requirements like problem statement, objectives, scope and significance have been identified. This project successfully designed system components including classification Model which implemented using SVM for emotion classification. Use case diagram that visualises system interactions and user roles. Flowchart diagram that outlined system workflow for clarity. User interface that designed to present classification results in a user-friendly manner.

So, the classification model was built using Support Vector Machine (SVM) which a widely used and effective algorithm for text classification. To enhance feature representation, Word2Vec SkipGram was implemented for feature extraction, enabling the model to capture contextual relationships between words and understand the semantic connections within crime-related news.

Furthermore, the project aims to classify six basic emotions, which are fear, sadness, joy, surprise, disgust and anger, but causing bias in predict of fear, sadness and anger. Therefore, this project should focus on expanding the dataset to include a broader range of news sources, languages, and integrating advanced natural language processing techniques like transformer-based models such as BERT could significantly improve the model's ability.

REFERENCES

Akimoto, T. (2020). Cogmic space for narrative-based world representation. Cognitive Systems Research, 65, 167–183. https://doi.org/10.1016/j.cogsys.2020.10.005

Alrasheedy, M. N., Muniyandi, R. C., & Fauzi, F. (2022). Text-Based Emotion Detection and Applications: A Literature review. *2022 International Conference on Cyber Resilience (ICCR)*, 1–9. https://doi.org/10.1109/iccr56254.2022.9995902

Andersen, K., Bjarnøe, C., Albæk, E., & De Vreese, C. H. (2016). How news type matters. *Journal of Media Psychology Theories Methods and Applications*, *28*(3), 111–122.

https://doi.org/10.1027/1864-1105/a000201

Burggraaff, C., & Trilling, D. (2017). Through a different gate: An automated content analysis of how online news and print news differ. *Journalism*, *21*(1), 112–129. https://doi.org/10.1177/1464884917716699

Christ, L., Amiriparian, S., Milling, M., Aslan, I., & Schuller, B. W. (2022, December 21). *Automatic emotion modelling in written stories*. Retrieved February 10, 2025, from http://arxiv.org/abs/2212.11382

Kennedy, A., Kazantseva, A., Inkpen, D., & Szpakowicz, S. (2012). Getting Emotional about News Summarization. In *Lecture notes in computer science* (pp. 121–132). https://doi.org/10.1007/978-3-642-30353-1_11

Li, M., Wang, D., Lu, Q., & Long, Y. (2016). Event Based Emotion Classification for News Articles. *Computing Department, the Hong Kong Polytechnic University, Hung Hom, Hong Kong*. Retrieved February 10, 2025, from https://www.researchgate.net/publication/317256374_Event_Based_Emotion_Classification_for_News_Articles

Takatsu, H., Ando, R., Matsuyama, Y., & Kobayashi, T. (2020). Sentiment analysis for emotional speech synthesis in a news dialogue system. *Proceedings of the 17th International Conference on Computational Linguistics* -. https://doi.org/10.18653/v1/2020.coling-main.440

Mohammad, S. M. (2020, March). *NRC affect intensity lexicon*. Saif M. Mohammad. Retrieve February 17, 2025, from https://saifmohammad.com/WebPages/AffectIntensity.htm

Nasir, A. F. A., Nee, E. S., Choong, C. S., Ghani, A. S. A., Majeed, A. P. P. A., Adam, A., & Furqan, M. (2020). Text-based emotion prediction system using machine learning approach. *IOP Conference Series Materials Science and Engineering*, *769*(1), 012022. https://doi.org/10.1088/1757-899x/769/1/012022

New Straits Times (NST Online) | Malaysia News & World Updates. (n.d.). NST Online. Retrieve February 17, 2025, https://www.nst.com.my/

Plutchik, R. (2001). The Nature of Emotions. *American Scientist*, *89*(4), 344–350. Retrieved February 10, 2025, from http://www.jstor.org/stable/27857503

Pólya, T., & Csertő, I. (2023). Emotion recognition based on the structure of narratives. Electronics, 12(4), 919. https://doi.org/10.3390/electronics12040919

The Star Online | Malaysia, business, sports, lifestyle and video news. (2019, July 29). Retrieve February 17, 2025, https://www.thestar.com.my/

TheSun | Malaysia News: National, world, viral & Free ePaper. (n.d.). thesun.my. Retrieve February 17, 2025, https://thesun.my/