Unveiling Mood Classifications in Malaysia: Analysing Code-Mixed Twitter Data for Emotional Expression

Latifah Abd Latib, Universiti Selangor, Malaysia
Hema Subramaniam, Universiti Malaya, Malaysia
Affezah Ali, Taylor's University, Malaysia
Siti Khadijah Ramli, Universiti Selangor, Malaysia

The Barcelona Conference on Arts, Media & Culture 2023
Official Conference Proceedings

Abstract

The fast rise of social media platforms has given academics unparalleled access to user-generated data, allowing for large-scale studies of public attitudes and moods. In a nation with such a rich culture as Malaysia, it is typical to see tweets written in Malay, local slang, and English. This language variation makes it more difficult to analyze emotions, especially given the need for labeled data required for supervised learning approaches. This study examines and categorizes Malaysians' mood expressions, mostly using code-mixing techniques discovered on Twitter. The study uses the Jupyter Notebook application to visualize and analyze a dataset comprising 2184 out of 2190 Twitter tweets after data pre-processing. The NRCLex Affect lexicon is used for both data analysis and emotion classification. The analysis reveals that approximately 50.9% of Twitter users were likely to express happiness, followed by 19.3% expressing trust, 10.9% expressing fear, 13.1% expressing sadness, 3.1% expressing anger, and 2.7% expressing surprise. The results are promising, as a relatively high level of accuracy was achieved even with a small initial labeled dataset. This outcome is significant when labeled datasets for emotion analysis are limited. Additionally, the research provides real-time analysis of emotions. The successful classification of mood expression in code-mixed tweets provides insights into Malaysians' emotional states, contributing to a deeper understanding of public sentiment. Understanding the prevailing mood is valuable in gauging public opinion, assessing social trends, and informing decision-making processes at both individual and societal levels.

Keywords: User-Generated Data, Language, Emotion Classification, Code-Mixing, Emotion Analysis

iafor
The International Academic Forum
www.iafor.org

Introduction

Social media is the most popular venue for sharing many facets of life, including emotions and feelings. With the rise of social media, emotions can now be expressed in writing, including status updates and posts that use punctuation, in addition to verbally and visually. The psychological and physiological phenomenon of emotion has multiple dimensions, including physiological changes, subjective emotions, cognitive processes, and behavioral reactions. It appears as a complicated state brought on by internal or external stimuli, profoundly affecting a person's perceptions and interactions with their environment (Brandy, 2013).

The primary emotions are contentment, trust, fear, surprise, sadness, disgust, wrath, and anticipation (Plutchik, 2003). These basic emotions are the foundation for many emotional experiences and are essential to how people explore and react to different circumstances throughout their lifetimes. An intriguing understanding of how people utilize online platforms to express their feelings and interact with others has emerged from research on emotional expression in social media. Social media status updates reveal a similar emotional profile to one's general emotional life across various emotions (Panger, 2017). Social media like Facebook and Twitter exhibit higher arousal levels than individuals' daily emotional experiences. Facebook posts were found to tend to be more positive expressions on average, while tweets tend to be more negative expressions (Panger, 2017).

Additionally, recent studies have shown the emotional contagion phenomenon, which shows how a user's feelings can affect those of other users in their social network. It could lead to the spread of both positive and negative sentiments on social media platforms (Lu & Hong, 2022). This occurrence emphasizes the important effects of revealing personal emotions online to other users, especially on social media platforms where both textual and visual content can be shared. Emotional expression is impacted by cultural norms and beliefs that emphasize group harmony and cohesion more than individual expression in collectivist countries like Malaysia. Open expressions of intense emotions may be viewed as disruptive or improper in a social setting; hence emotions are frequently expressed subtly and restrainedly. As a result, emotional expression is frequently more guarded and subdued out of a desire to preserve one's "face" or good reputation. People might be reluctant to publicly express unpleasant feelings like anger or disappointment, preserve their credibility, and maintain positive relationships with others (Mesquita, 2001). Social media has become increasingly popular for expression since its anonymity enables people to do so without being constrained by face-to-face encounters. Collectivist societies are more likely to use social media more frequently and openly express their emotions online (Alsaleh et al., 2019).

Social Media and User-Generated Data (UGD)

Social media now dominates our daily lives. It has both positive and negative outcomes. Additionally, social media can enable us to maintain relationships with loved ones, keep up with current affairs, support organizations and causes, and give people more influence. For example, social media can make it easy to stay in touch with people who live far away, share photos and videos, and chat live. It can also be a great way to stay informed about current events by following news organizations, politicians, and other thought leaders. Additionally, social media can promote businesses and causes by reaching a large audience with your message and building relationships with potential customers and supporters.

The vast attraction of social media platforms transcends geographical and cultural boundaries. Social media today significantly influences everyone's daily lives by promoting relationships, information sharing, and communication. In 2023, there will be 4,9 billion social media users worldwide. There are now 4.91 billion users of social media worldwide, which is a record high. By 2027, this figure is projected to rise to 5.85 billion. (Global Social Network, 2023). With 2,9 million monthly active users across the globe, Facebook remains the most popular social media platform. In line with global trends, Malaysia has enthusiastically embraced social media. It has a thriving digital landscape with a high internet penetration rate, contributing to social media platforms' pervasive adoption. 78.5 per cent of Malaysians were active social media users in January 2023. Compared to 2022, when roughly 91.7% of Malaysia's population used social media, this marked a fall of 13.2% (Global Social Network, 2023). Facebook continues to be one of the most popular social networking, communication, and content-sharing platforms among Malaysians. WhatsApp, a Facebook-acquired messaging application, is also wildly popular in Malaysia, functioning as a primary means of communication for many users. Instagram is widely used by MalWith a sizable user base, Malaysians, especially younger demographics, as a platform for sharing photos and stories, and interactions (2023, Statista). With a sizable user base, Twitter is a platform for public discussions, news updates, and prominent topics in Malaysia. Overall, the popularity of social media platforms in Malaysia parallels the global trend, reflecting the digital transformation of society and the increasing influence of social media on communication, information exchange, and social interactions. As technology evolves, social media will likely remain a prominent force in Malaysia and around the globe, requiring users and stakeholders to address the opportunities and challenges it presents.

Social Media and User-Generated Data (UGD)

User-generated data (UGD) is any text, data, or action created by online digital systems such as social media users. This content is published and disseminated by the user through independent channels and can have an expressive or communicative effect, either on its own or when combined with other contributions from the same or other sources. UGD can be used to improve customer insights by providing businesses with valuable information about their customers' needs, wants, and behaviours. This data can be utilized to enhance goods and services, create fresh advertising campaigns, and more effectively target consumers. Businesses can use social media, for instance, to survey client requirements and preferences or gather feedback on new product prototypes.

By enabling users to share their good interactions with a brand, UGD can also raise brand awareness. It might be a successful tactic for bringing in new customers and encouraging loyalty among existing ones. For instance, businesses encourage customers to review their products or services or to publish images and videos of their purchases on social media. Although there may be opportunities, UGD could also present hazards. The greatest danger is data privacy. Businesses must use caution while gathering and using UGD because it frequently contains personal data about specific people. Businesses risk legal consequences and reputational damage if they fail to protect the privacy of UGD. Misinformation is a danger connected to UGD. UGD can be used to disseminate inaccurate or deceptive information, which could be detrimental to society. For example, businesses could use UGD to create fake reviews or to spread false information about their competitors. Businesses need to be aware of the potential for misinformation when using UGD, and they need to take steps to verify the accuracy of this data. However, UGD can also be a tool for online bullying. When someone bullies someone online, generally by sending scary or threatening messages, this is known as cyberbullying. Businesses need to be aware of the possibility of cyberbullying on their platforms because it can have catastrophic effects on victims. UGD via social media platforms may be a potent tool for businesses. Nevertheless, being conscious of such risks and taking precautions to reduce them is crucial.

Code-Switching

The linguistic landscape of Malaysia is characterized by its multi-racial composition, which has endowed Malaysians with the ability to converse in at least two distinct languages. This phenomenon of bilingualism and its concomitant linguistic flexibility have fostered the natural development of code-switching in both spoken and written discourse among Malaysians. The practice of code-switching is particularly pronounced in Malaysia due to its multicultural milieu. Recent research by Tan et al. (2020) has revealed that this linguistic phenomenon extends to online textual communication, further emphasizing its prevalence and importance in the Malaysian sociolinguistic context. Moreover, Malaysia's language policy, driven by its multiethnic and multilingual population, sets the country apart. Bahasa Malaysia, the official language, is complemented by English, which is mandated for educational purposes and serves as the second language of the nation.

The complex interplay of languages, such as Malay-English code-switching, is an area of significant interest, with its widespread use in formal and informal settings making Malaysia an intriguing case for the study of code-switching (Zulkifli & Tengku Mahadi, 2020). This linguistic landscape, as indicated by Treffers-Daller et al. (2022) in the available literature, presents a rich tapestry of language switching patterns, even in formal legal contexts. The advent of technology and the proliferation of social media platforms have further broadened the scope for linguistic communication, enabling individuals to engage in discourse on a more extensive scale.

A widespread misunderstanding is that Malaysian English is an "improper" rendition of the English language that makes communication difficult, particularly with native English speakers (Rahim, 2022). It unites Malaysia's multilingual speakers. Malaysian English is a strong, widely used variation of English throughout the nation, regardless of its format—spoken or written, formal or informal. Most Malaysians who speak English as a second language use Malaysian English, popularly known as "Manglish," a regionalized variety of languages with unique elements that vary in meaning, sound, and structure. It is acknowledged as a fresh variation of English, like the accepted language varieties spoken in Singapore, India, and the Philippines. With a distinctive blend of regional terms, pronunciation, intonation, and liberal use of the particle "lah" or "la," its informal variations are more frequently heard in casual discussions, store exchanges, and street interactions. This style is also called "rojak English," after the well-known Malaysian dish "rojak," renowned for its various ingredients.

In the part of this study (Zabha et al., 2019), a large lexicon classifier was used to experiment with sentiment analysis on Twitter users in Malaysia, and the accuracy of the results was assessed. It was determined whether the words in the Malay and English lexicons should be classified as positive or negative. For the English lexicon, the most frequent pair of nouns or pair of noun phrases that come at the beginning of a feeling is taken as samples of opinions, whether negative or positive points of view, and the most frequent sample in a positive opinion will be deemed a "positive word" and vice versa. For the Malay lexicon, Wordnet is used to rate the terms based on their meaning and synonyms, and the Naïve Bayes technique is then used to double-check the correctness of the Wordnet-awarded points. This study suggested building a database of positive and negative terminology using a blend of English and Malay and using R to identify whether the attitude was positive or negative. The project aimed to create a cross-lingual sentiment analysis utilizing a lexicon-based methodology based on prior research. The technology combines the lexicons of two languages. After that, Twitter data was gathered, and a graph was used to display the results. The outcomes demonstrated that the classifier could extract the attitudes. This study is important for businesses and governments to understand how people feel about social media, especially in Malay-speaking areas. The other study (Kamble & Joshi, 2018) updates the best practices for identifying hate speech in tweets with English and Hindi coding. They utilize domain-specific embeddings to compare three prevalent deep-learning models. Employing a benchmark dataset of tweets with English and Hindi code combined, we conducted experiments and found that employing domain-specific embeddings results in a better representation of target groups. They further demonstrate that our models improve the F-score by roughly 12% compared to earlier work that used statistical classifiers.

The pre-processing methods used to normalize the noisy text, the most common performance metrics for Malay Sentiment Analysis (SA), and the difficulties for Malay SA have yet to be examined (Abu Bakar et al., 2020). The analysis of emotions in these code-mixed tweets presents certain difficulties. First, the language has a mixture of official and casual single-word terminology and multi-word expressions (MWE). Second, non-standard sentence structures are used since the constituent languages of the code-mixed text have diverse vocabularies and grammatical features. Thirdly, the number of labeled code-mixed datasets that can be used as training datasets is limited. Fourth, there are only so many resources for mood and feeling in languages other than English, especially for languages with several writing systems (Tan et al., 2020).

Our overall goal is to categorize each text message (tweet) into one of several emotion classes. Existing approaches can be divided into lexical methods and machine learning approaches. The dataset used for the experiments was the publicly available Malay and English emotion dataset, a collection of Twitter data organized into six files according to the emotion's happiness, sadness, trust, fear, anger, surprise, disgust, and joy. The tweets in this dataset are mostly in Malay and Malaysian slang and code-mixed Malay-English text. Details of the dataset's composition according to the different emotion classes are shown in **Table 1** and **Figure 1 shows** the rating distribution according to the different emotion classes.

Table 1: *Dataset's composition according to the different emotion classes*

Emotion	No of Tweets
positive	1112
Trust	421
Fear	273
Negative	188
Anticipation	94
Anger	67
Surprise	60
Joy	2
Sadness	2
disgust	1
Total	2184

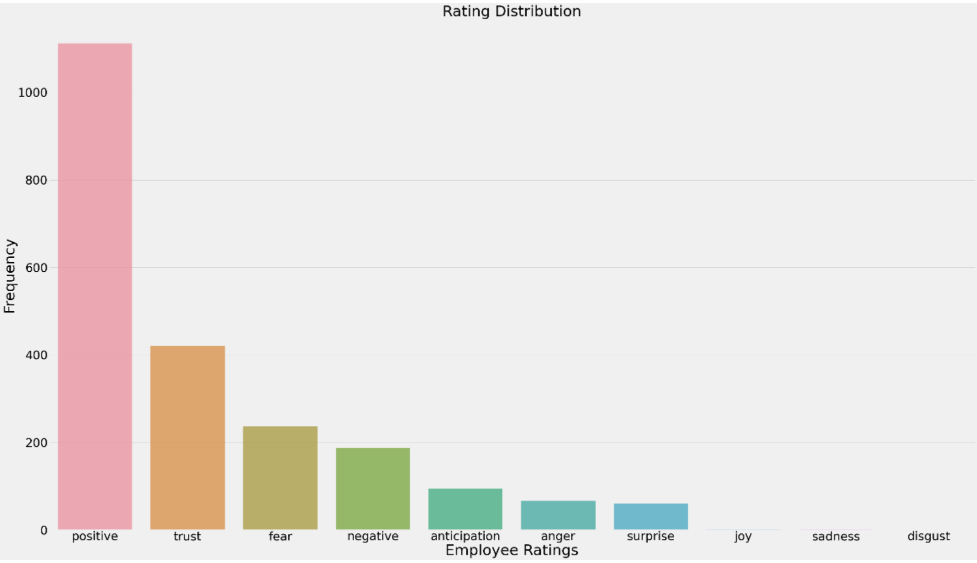

Figure 1: Rating distribution according to the different emotion classes

Conclusion

Based on the results of the data preprocessing techniques and the classification algorithm, the following conclusions can be drawn the emotion analysis of the Twitter dataset reveals the prevalence of different emotions among the collected tweets. Most tweets (approximately 1587) were classified as expressing happiness, indicating a positive emotional tone in the data. Fear and anger were less prevalent emotions, with 252 and 194 tweets classified under these categories, respectively. Sadness was the least prevalent emotion, with 151 tweets categorized as expressing sadness.

The application of clustering techniques in the classification algorithm has significantly influenced the distribution of emotional classes. The increases in the number of instances classified into each emotional class demonstrate the effectiveness of these techniques in improving the accuracy of emotion analysis. Clustering has likely helped capture underlying patterns and relationships in the data, leading to better emotion classification. From the emotion analysis insights, understanding the emotional content of Twitter data can provide valuable insights into the sentiments and feelings of users. The dominance of happiness-related tweets may indicate a generally positive sentiment among Twitter users in the analyzed dataset. On the other hand, the presence of fear, anger, and sadness-related tweets highlights that diverse emotional expressions are present on the social media platform. Emotional analysis can track public mood and opinions on various topics, including political ideologies, social gatherings, and prominent figures. Researchers can use this data to comprehend public opinion and make fact-based conclusions. In mental health support systems, emotion analysis can identify indicators of emotional discomfort or depression by examining user text input. It can assist in locating those who require more assistance and intervention.

In conclusion, the emotion analysis of the Twitter dataset demonstrates the distribution of different emotions expressed by users. The application of clustering techniques has contributed to significant improvements in emotion classification accuracy. These findings can offer valuable insights into user sentiments and have various applications in social media monitoring and market research. However, ongoing research and validation are necessary to continually refine and enhance the emotion analysis process.

Acknowledgments

This research was financially supported by the Ministry of Higher Education (MOHE) Malaysia under the Fundamental Research Gant Scheme (FRGS) Universiti Selangor (FRGS/1/2020/SS0/UNISEL/02/2).

References

Abd Rahman, A. A., & Abdul Razak, F. H. (2019). Social Media Addiction Towards Young Adults Emotion. Journal Of Media And Information Warfare, 12(2), 1–15.

Abu Bakar, M. F. R., 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. https://doi.org/10.1109/ACCESS.2020.2968955

Alsaleh, D. A., Elliott, M. T., Fu, F. Q., & Thakur, R. (2019). Cross-cultural differences in the adoption of social media. Journal of Research in Interactive Marketing, 13(1), 119–140. https://doi.org/10.1108/JRIM-10-2017-0092

Brady, M. (2013). Emotional insight: The epistemic role of emotional experience. Oxford University Press.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. Nebraska Symposium on Motivation, pp. 19, 207–283.

Ekman, P. (1992). Are there basic emotions? Psychol. Rev. 99, 550–553. doi:10.1037/0033-295X.99.3.550

Global Social Network. (2023). Global Social Media Statistics. Retrieved from https://datareportal.com/social-media-users

Kamble, S., & Joshi, A. (2018). Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models. http://arxiv.org/abs/1811.05145

Lu, D., & Hong, D. (2022). Emotional Contagion: Research on the Influencing Factors of Social Media Users' Negative Emotional Communication During the COVID-19 Pandemic. Frontiers in Psychology, 13, 931835. https://doi.org/10.3389/fpsyg.2022.931835

Mesquita, B. (2001). Emotions in collectivist and individualist contexts. Journal of Personality and Social Psychology, 80(1), 68–74. DOI:10.1037/0022-3514.80.1.68

Mohd Yuswardi, P. N. A. S., & Ahmad, N. A. (2023). Sentiment Analysis of Malaysian Citizen's Emotion towards Cyberbullying on Twitter. International Journal of Academic Research in Business and Social Sciences, 13(4), 769–780. https://doi.org/10.6007/ijarbss/v13-i4/16777

Ng,Y.M.M & Taneja, H. (2023). Web use remains highly regional even in the age of global platform monopolies. PLoS ONE 18(1): e0278594. https://doi.org/10.1371/journal.pone.0278594

Ortony, A. (2021). Are All “Basic Emotions” Emotions? A Problem for the (Basic) Emotions Construct. Perspectives on Psychological Science, 17(2), 174569162098541. DOI:10.1177/1745691620985415

Panger, G. T. (2017). Emotion in Social Media. UC Berkeley Electronic Theses and Dissertations. Retrieved from https://escholarship.org/uc/item/1h97773d

Plutchik, R. (2003). Emotions & Life. Perspectives From Psychology, Biology, and Evolution. Washington, DC: American Psychological Association.

Rahim, H. A. (2022). Malaysian English is not mangled; it's unifying. Newhub 360, 1–1. https://360info.org/malaysian-english-is-not-mangled-its-unifying/

Razak, C. S. A., Hamid, S. H. A., Meon, H., Subramaniam, H. A., & Anuar, N. B. (2021). Two-Step Model for Emotion Detection on Twitter Users: A Covid-19 Case Study in Malaysia. Malaysian Journal of Computer Science, 34(4), 374–388. https://doi.org/10.22452/mjcs.vol34no4.4

Statista. (2023). Demographics of Instagram users in Malaysia as of June 2023 by age group. Retrieved from https://www.statista.com/statistics/1399780/malaysia-demographics-of-instagram-users-by-age-group/

Suhasini, M., & Badugu, S. (2018). Two-Step Approach for Emotion Detection on Twitter Data. International Journal of Computer Applications, 179(53), 12–19. https://doi.org/10.5120/ijca2018917350

Tan, K. S. N., Lim, T. M., & Lim, Y. M. (2020). Emotion analysis using self-training on Malaysian code-mixed Twitter data. Proceedings of the 13th IADIS International Conference ICT, Society and Human Beings 2020, ICT 2020 and Proceedings of the 6th IADIS International Conference Connected Smart Cities 2020, CSC 2020 and Proceedings of the 17th IADIS International Conference, 181–188. https://doi.org/10.33965/ict_csc_wbc_2020_2020081022

Treffers-Daller, J., Majid, S., Thai, Y. N., & Flynn, N. (2022). Explaining the Diversity in Malay-English Code-Switching Patterns: The Contribution of Typological Similarity and Bilingual Optimization Strategies. Languages, 7(4). https://doi.org/10.3390/languages7040299

Warren, G., Schertler, E., & Bull, P. (2009). Detecting Deception from Emotional and Unemotional Cues. Journal of Nonverbal Behavior, 33(1), 59-69. DOI:10.1007/s10919-008-0057-7

Zabha, N. I., Ayop, Z., Anawar, S., Hamid, E., & Abidin, Z. Z. (2019). Developing cross-lingual sentiment analysis of Malay Twitter data using a lexicon-based approach. International Journal of Advanced Computer Science and Applications, 10(1), 346–351. https://doi.org/10.14569/IJACSA.2019.0100146

Zulkifli, Z. I., & Tengku Mahadi, T.-S. (2020). Reasons to Code-Switch : A Case Study of Malaysian Twitter Users. June, 35–43.

Contact email: latifah@unisel.edu.my