International Journal of Innovation and Multidisciplinary Research
DOI.ORG/10.67505/IJIAMR.v6.i1.61008
Vol. 6 (1), 2026
E-ISSN: 2583-4452

Sentiment Analysis in Emoji-Rich Text: A Comprehensive Survey of Techniques, Datasets, Challenges and Research Directions

Prof. Anamika Gupta[1], Sugandha Kaur[2*], Vanshika Goyal[3], Aradhya Jain[4] and Ravika Singh[5]
[1,3,4,5] Shaheed Sukhdev College of Business Studies, University of Delhi, Delhi
[2] Mata Sundri College for Women, University of Delhi, Delhi
[*] Corresponding Author ✉ sugandhakaur@ms.du.ac.in

ABSTRACT

Visual digital communication is advancing substantially, with emoji-rich text, making it quite challenging for the analysis of sentiments from such text. In the paper, a comprehensive overview is provided for the methodologies for sentiment analysis in emoji-rich text. In this study, the landscape of available datasets curated for this text is reviewed. Their relevance, limitations, and linguistic diversity is also studied. Furthermore, a comprehensive review of the preprocessing techniques and the machine learning models utilized is also presented. A detailed study is conducted to highlight the specialized processing techniques used for the interpretation and handling of emojis, such as emoji embeddings and ranking systems. Further, research gaps are highlighted, including a multilingual and multimodal approach, and non-availability of high quality datasets. The paper concludes with a discussion on future work advocating for context-sensitive models and the use of generative AI for the development of emoji-aware pre-trained models. This work will serve as a comprehensive reference point for researchers planning to advance sentiment analysis of the emoji-driven digital world.

Keywords : *Emoji-rich, Sentiment Analysis, Emoji Embedding, Context-sensitive*

1 Introduction

Sentiment analysis is a field that blends Machine Learning (ML) and Natural Language Processing (NLP), enabling computers to understand and interpret the emotions expressed behind written text. The sentiment behind the text could be classified as positive, negative, or neutral Ayvaz and Shiha (2017); Yoo and Rayz (2021). With the advent of technology, a huge amount of data is generated on a daily basis through social networks, online reviews, discussion forums, and news articles. Manual analysis and interpretation of this amount of data is practically impossible Alfreihat et al. (2024). The process of sentiment analysis can help in such scenarios by extracting valuable insights from huge volumes of data, which in turn aids in detecting feelings, opinions, or attitudes of the author toward a specific topic, product, or event. It also helps to monitor public sentiment in real time, allowing companies, governments, and researchers to respond quickly to emerging trends, issues, or concerns Khan et al. (2025). In a nutshell, sentiment analysis imparts stakeholders a clearer understanding of how their audiences feel, helping them in thoughtfully addressing their feedback and concerns.

Sentiment analysis is not only confined with the written text, but it also takes into account emojis and emoticons, which play a crucial role in the present digital communication eraHakami et al. (2022); Yuan et al. (2022). An emojis is a tiny graphical symbol or icon that represents ideas, objects or emotions Holthoff (2020); Kralj Novak et al. (2015). On the other hand, an emoticon is a combination of letters and punctuation marks arranged to make a facial expression Hogenboom et al. (2013); Kralj Novak et al. (2015); Srisankar and Lochanambal (2021); Ullah et al. (2020). Emoticons were used even before emojis came. Emojis and emoticons add an extra layer of special emotional meaning to the plain text messages Karthikeya et al. (2023). In day-to-day offline communications, we rely more on body language, intonations, and facial expressions to understand how the other person feels. But in digital communication, these non-verbal cues are missing. That's where, both emojis and emoticons come into the picture and aids in a better understanding of the intent of the communication; thereby reducing the chances of being misunderstood Hogenboom et al. (2013); Santos et al. (2023).

Received on: 11 Jun 2026 | Accepted on: 01 Jul 2026 | Published Online on: 25 July 2026

Because of this, both emojis and emoticons play a key role in sentiment analysis by giving direct clues about how someone feels. These symbols provide extra emotional context that helps computers better understand the realsentiment of a message. Without these symbols, it can be easy to misinterpret the tone of the messageAyvaz and Shiha (2017); Lou et al. (2020).

In this paper, a comprehensive literature review is conducted for the work done in the area of sentiment analysis on emoji-rich text. The research objectives of this paper are mentioned below:

Research Objectives

1. RQ1: What are the characteristics of existing datasets, like size, scope, annotations used, web-scrapped or downloaded etc?
2. RQ2: What are the preprocessing, feature extraction and machine learning techniques used to handle text containing emojis?
3. RQ3: What are the specialized methods for handling emojis?
4. RQ4: What are the key challenges and research gaps in sentiment analysis of emoji-rich text?

Organization of the Paper

The paper is divided into eleven sections. Section 2 outlines the systematic survey methodology. Section 3 explores various datasets used by researchers. Machine learning pipeline including the preprocessing and feature engineering is discussed in sections 4, 5 and 6. Sections 7 and 8 highlights the analysis done on the research trends on quantitative and qualitative basis. Section 9 discusses various challenges of handling emojis. Section 10 highlights the research gaps and synthesis of our research questions RQs. Finally, section 11 concludes our study and emphasizes on the future work in this domain.

2. Survey Methodology

The following section describes the methodology used in this review paper:

- **Databases and Sources:** Literature was sourced from major academic repositories like IEEE Xplore, ACM Digital Library and ScienceDirect to ensure high-quality peer-reviewed content.
- **Search Strings:** The search string used is: *("sentiment analysis" OR "emotion recognition") AND ("emoji" OR "emoticons") AND ("ma-chine learning" OR "deep learning" OR "NLP" OR "Natural Langugae Processing")*
- **Selection Criteria:** Papers were included based on their direct relevance to the research objectives (RQ1-RQ4). Specifically, we prioritized studies that introduced unique emoji-rich datasets, novel preprocessing pipelines, or specialized emoji embedding and modeling architectures. Studies that treated emojis merely as noise to be removed without specialized interpretation were generally excluded to focus on emoji-aware methodologies.
- **Time Window:** The reviewed literature spans from the period of 2010 to 2025; showcasing the evolution from initial emoticon based analysis to modern AI generative approaches.
- **Study Selection Counts:** A total of 28 primary research papers were selected for in-depth analysis and comparative benchmarking (Note: While 28 papers were included for analysis, approximately 150 papers were initially screened based on titles and abstracts to reach this final selection).

This survey establishes emojis as the primary carriers of sentiments due to their contextual role assignments and multidimensional emotional design. Existing reviews lacks the systematic rigor and linguistic diversity of this review, which is conducted in 19 languages and 32 countries. This review also attempts to synthesize empirical performance benchmarks and evaluates the application of multimodal generative AI for zero-shot cross-lingual analysis.

3. Characteristics of the Datasets

As with all attempts at natural language processing, the success of sentiment analysis that incorporates emojis depends upon the use of an appropriate dataset. As such, the characteristics of the available datasets for sentiment analysis that incorporate emojis were reviewed in this stage of the survey to determine their relevant features. Such features include the origin of the dataset, the size of the dataset, the linguistic diversity of the dataset, and the diversity of emojis included within those datasets.These dataset characteristics directly determine the robustness, accuracy, and generalizability.

This section compares and contrasts these dataset characteristics with respect to several influential papers in depth for further understanding. The comparison framework focuses on: the nature of the data source, the dataset's scale, its acquisition method (ready-made download vs. researcher-scraped), the extent of emoji inclusion for modelling and linguistic diversity of the datasets used.

3.1 Dataset Sources

The analysis of existing research reveals a distinct predominance of social media platforms as the primary source for datasets used in emoji and emoticon sentiment analysis. Twitter emerges as the overwhelmingly dominant platform, explicitly cited in majority of the listed studies Ayvaz and Shiha (2017); et al. (2020); Godard and Holtzman (2022); Jagadishwari et al. (2021); Jona et al. (2022); Karthikeya et al. (2023); Kralj Novak et al. (2015); Mona M. Abd Elsalam (2022); Ullah et al. (2020); Vashisht and Jailia (2020); Yoo and Rayz (2021); Yuan et al. (2022). Its prevalence is attributed to its public nature, high volume of emoji usage, and accessible APIs (e.g., Twitter Stream API, Twitter Stream Grab project) facilitating data collection, whether through direct downloads or web scraping. Several Twitter datasets focus on specific events or timeframes, such as COVID 19, New Years Eve , Istanbul Attack , or US Airline sentimentKhan et al. (2025); Ullah et al. (2020). Kaggle Mona M. Abd Elsalam (2022); Yoo and Rayz (2021) and GitHub are other curated data repositories which are readily used by the researchers.

3.2 Size and Scope of Datasets

The scale and extent of the datasets used for emoji sentiment analysis differ greatly. While, on the one hand, there are narrowly scoped datasets like the few images Alfreihat et al. (2024), there also exist massive ones such as the Twitter dataset including millions of tweets Yin et al. (2023); Ullah et al. (2020). This can be seen from the fact that the amount of data is even reflected in the Twitter Stream Grab dataset amounting to 3.8 TB of storage space Tang et al. (2024). In regards to scope, most of the datasets are targeted at a very particular context or even, for example the COVID 19 dataset Zhao et al. (2024), terrorism Vashisht and Jailia (2020), or airline sentiment Santos et al. (2023); Wang and Castanon (2019). Although many of the datasets are generated from only one platform, namely Twitter Mona M. Abd Elsalam (2022); Yin et al. (2023); Wang and Castanon (2019); Ullah et al. (2020); Toet et al. (2018), the remaining ones try to provide a broad scope using data from various sources, such as -various social media platforms Hakami et al. (2022); Khan et al. (2025)-, Facebook or gaming forums.

In any case, the high degree of variability presents an ongoing problem, since researchers have to consider the trade-off between the possibilities presented by big data, narrow-scoped studies and the feasibility and effectiveness of the approach to emoji sentiment analysis.

3.3 Downloaded vs. Web Scraped Data

The method used for the collection of the dataset is often categorized as downloaded datasets or scrapped datasets. For example, many datasets are *downloaded* from their platforms. These downloaded datasets can include anything from the initial release of the sentiment analysis dataset, to datasets that were collected using the social media platforms APIs, and from the various research data sharing platforms. Furthermore, many sentiment analysis datasets are also *scraped* from social media platforms. As with downloaded datasets, scraped datasets can also include data from various social media platforms, such as Twitter, Facebook, gaming forums, or social media platforms in general. Finally, as with downloaded datasets, the data that is scraped can also be targeted towards specific topics of interest. For instance, sentiment analysis research has used scraped data to focus upon social media discussions of topics such as the COVID 19 pandemic or the Istanbul attack.

The studies Alfreihat et al. (2024); Hakami et al. (2022); Jagadishwari et al. (2021); Khan et al. (2025); Mona M. Abd Elsalam (2022); Santos et al. (2023); Tang et al. (2024); Toet et al. (2018); Yin et al. (2023); Yoo and Rayz (2021) used downloaded datasets while the studies Alawneh et al. (2024); Ayvaz and Shiha (2017); et al. (2020); Godard and Holtzman (2022); Hogenboom et al. (2013); Holthoff (2020); Jona et al. (2022); Karthikeya et al. (2023); Kralj Novak et al. (2015); Lou et al. (2020); Srisankar and Lochanambal (2021); Ullah et al. (2020); Vashisht and Jailia (2020); Wang and Castanon (2019); Yoo and Rayz (2021); Yuan et al. (2022); Zhao et al. (2024) used web scraped datasets. The study Yoo and Rayz (2021) used a mix of both approaches for datasets, three of its datasets are downloaded, and one is web scraped.

3.4 Emoji Coverage

Once researchers have their data (whether downloaded or scraped), they face a critical hurdle, i.e., deciding on which emojis shall be included in their study (coverage). These choices profoundly shape the understanding and usefulness of the model. Most studies report coverage between 3 and 21 percent, while several omitting exact figures. The studies reported that emoji coverage ranges from 2.8 percent in the dataset (Istanbul Attack tweets, 2017) of Kralj Novak et al. (2015) to 100 percent in Tang et al. (2024); Srisankar and Lochanambal (2021).

3.5 Linguistic Diversity

Research on emoji sentiment analysis uses datasets that span several major languages in the world, reflecting the global nature of digital communication. English is the most consistently represented language, appearing in dedicated studies or alongside other languages in multilingual datasets. It is explicitly covered in studies analyzing Portuguese and English tweets (e.g., elections and World Cup data) , Dutch and English social media messages , and Chinese and English microblogs. Arabic has emerged as a major focus in contemporary research, with multiple dedicated studies leveraging large Arabic datasets Alawneh et al. (2024); Alfreihat et al. (2024); Hakami et al. (2022). Portuguese features in targeted bilingual research alongside English . Chinese is prominently represented, particularly through large-scale collections of

Sina Weibo microblogs Lou et al. (2020); Zhao et al. (2024). The broadest linguistic scope comes from a large-scale multilingual study encompassing 13 distinct languages, analyzing tweets from diverse linguistic backgrounds. Dutch is also included, specifically analyzed in conjunction with English social media data.Table 1 summarizes the findings.

Table 1: *Comparison of Emoji-Rich Studies*

Ref	Year	Language	Platform/Source	Size	Label Set	Method	Emoji Statistics	Authentic Access Link
[1]	2024	Arabic	Kaggle	56,000	3-way	Scraped	912 Mapped Emojis	https://www.kaggle.com/datasets/mksaad/arabic-sentiment-twitter-corpus
[2]	2024	Arabic	GitHub	58,000	3-way	Downld.	3.82%	https://github.com/manaralfreihat/Emoji-Sentiment-Lexicon
[3]	2017	English	Twitter	101k/73k	3-way	Scraped	19.38%	http://www.independent.co.uk/news/world/europe/istanbul-nightclub-attack-shooting-turkey
[4]	2020	English	Twitter	Not stated	3-way	Scraped	Not stated	Twitter (Surikov et al.)
[5]	2022	English	Twitter API	3,000,000	3-way	Scraped	21 Emoji Symbols	https://www.frontiersin.org/articles/10.3389/fpsyg.2022.921388/full
[6]	2022	Arabic	Twitter	144,196	3-way	Downld.	Not stated	https://archive.org/details/twitterstream
[7]	2013	Dutch/Eng	Google/Twitter	2,080	3-way	Scraped	100%	Twitter / Google Discussion Groups
[8]	2020	English	Various SM	Not stated	3-way	Scraped	Not stated	Various Social Media Platforms
[9]	2021	English	Kaggle	3,000,000	3-way	Downld.	Not stated	https://www.kaggle.com/youben/twitter-sentiment
[10]	2022	English	Twitter/Forum	Not stated	3-way	Scraped	Not stated	Twitter / Gaming Forums
[11]	2023	English	Twitter API	Not stated	3-way	Scraped	Not stated	Twitter Stream API
[12]	2025	English	Kaggle	14,640	3-way	Downld.	6 Emoji Symbols	https://www.kaggle.com/crowdflower/twitter-airline-sentiment
[13]	2015	English	Twitter	1,600,000	3-way	Scraped	100 Emoji Symbols	Twitter (Kralj Novak et al.)
[14]	2020	Chinese/Eng	Sina Weibo	10,042	3-way	Scraped	Not stated	Sina Weibo Platform
[15]	2022	English	Kaggle	10,000	3-way	Downld.	Not stated	https://www.kaggle.com/imranzaman5202/starter-code-covid-19-sentiment-analysis
[16]	2017	Multiling.	Facebook/Tw.	460 KW	25 Emot.	Scraped	Not stated	Facebook / Twitter
[17]	2023	Port./Eng	Kaggle	84,115	3-way	Downld.	Not stated	https://www.kaggle.com/datasets/adilmar/twitter-pt
[18]	2024	Code-mixed	Zenodo	20,000	Multi-label	Downld.	Not stated	https://zenodo.org/record/3974927
[19]	2021	13 Langs	Twitter	1,600,000	3-way	Scraped	4 Emoji Symbols	Twitter (Srisankar et al.)
[20]	2024	English	Multi-Twitter	Multi-DS	3-way	Downld.	Not stated	Twitter (Tang et al.)
[21]	2018	N/A	FRIDa	60 Images	Emotion	Downld.	100%	FRIDa Image Database
[22]	2020	English	Twitter/Reviews	14,460	3-way	Scraped	5.06%	Twitter / Airline Reviews
[23]	2020	English	Various SM	Not stated	3-way	Scraped	Not stated	Various Social Media Sources
[24]	2019	English	Twitter API	13,000,000	3-way	Scraped	50 Emoji Symbols	https://uvaauas.figshare.com/articles/dataset/TwemojiDataset/5822100
[25]	2023	English	Twitter Grab	184,485,91	93-way	Downld.	Not stated	https://archive.org/details/twitterstream
[26]	2021	English	Kaggle/GitHub	1,800,000	3-way	Downld.	Not stated	Kaggle / GitHub / Twitter API
[27]	2022	English	Twitter	Not stated	3-way	Scraped	Not stated	Twitter (Yuan et al.)
[28]	2024	Chinese	Sina Weibo	6,100,000	3-way	Scraped	> 50 Emoji Symbols	Sina Weibo API

3.6 Ethical Considerations in Data Acquisition

Data acquisition through web-scraping or downloading from repositories has some serious ethical concerns. Firstly, the terms of service must be respected for the public platforms like Twitter, Kaggle, and Facebook, which are the prime sources of data collection. Next, the privacy of personal information must be protected by de-identifyingtext, such as removing personal details like usernames or geographic coordinates, so that no single person can be traced back through their unique use of emojis. Further, sharing of the data must be done with caution, by utilizing research licenses on platforms like Kaggle or sharing only post IDs rather than raw text. Lastly, fairness and representation must be addressed by addressing potential biases, which helps minimize the confounding effects of gender or culture in the results.

3.7 The Emoji-Aware Sentiment Analysis Pipeline

The main steps that are often taken to process emoji-rich text are generally divided into three main stages. This ensures that the raw symbols are translated into sentiment signals that can be interpreted by the machine.

3.8 Stage 1: Emoji Preprocessing and Normalization

The first step in this emoji-enabled process involves processing of raw symbols in preparation for modeling, identifying emojis that often affect sentiments through emphasis, clarification and reversing. Categorical filtering is an important part of this process, wherein researchers filter the emojis using categorical filters such that only facial expressions and gestures are selected, and other non-relevant information is stripped off, thus avoiding any form of cultural and gender bias during the analysis process. Substitution and encoding strategies are used to translate emojis into descriptive text to maintain consistency across diversified platforms. Furthermore, similar emojis are grouped using normalization techniques to minimize feature sparsity.

3.9 Stage 2: Feature Representation and Vectorization

The second stage in sentiment analysis that incorporates emojis is feature representation and vectorization of the emojis. More specifically, emojis are often represented via the use of either static sentiment values or dynamic emoji embeddings to represent each emoji within a text to be analyzed. Lexicon-based mapping is one of the static mapping techniques readily used, where in dictionaries like EmojiNet or Emoji Sentiment Lexicon (ESL) is used to map emojis with a polarity score. This score is calculated by finding the difference between the positive and negative polarity of the emoji. Furthermore, dynamic embeddings include the use of Emoji2Vec, DeepMoji, or EmoGraph2vec. These map each emoji to a dynamic high-dimensional vector that represents the emoji's sentiment towards text.

3.10 Stage 3: Classification Architectures

The final and last main stage for sentiment analysis that incorporates emojis is the use of various classification models (Machine Learning or Deep Learning) to classify the text and determine the sentiment of that text.

4. Preprocessing of Features

Data cleaning and preprocessing, a crucial step in text mining, removes noise from data, reduces the vocabulary size, and prepares raw data for modeling. A clean and standardized input to machine learning models makes them efficient and more accurate. Some of the commonly used data cleaning and preprocessing techniques are mentioned in Table 2. Lowercase, usually the first step, standardizes the text by converting all the characters to lowercase, tokenization Alawneh et al. (2024); Alfreihat et al. (2024); Ayvaz and Shiha (2017); Hogenboom et al. (2013); Jagadishwari et al. (2021); Karthikeya et al. (2023); Khan et al. (2025); Kralj Novak et al. (2015); Lou et al. (2020); Mona M. Abd Elsalam (2022); Rahman et al. (2017); Singh et al. (2024); Srisankar and Lochanambal (2021); Tang et al. (2024); Ullah et al. (2020); Vashisht and Jailia (2020); Yin et al. (2023); Yoo and Rayz (2021); Zhao et al. (2024) splits the text into smaller units which can be words, sentences, paragraphs, documents, etc. Non-informative symbols like punctuation marks and special characters are handled mostly by removing them. Stop words, Alfreihat et al. (2024); Ayvaz and Shiha (2017); Jagadishwari et al. (2021); Jona et al. (2022); Khan et al. (2025); Kralj Novak et al. (2015); Lou et al. (2020); Rahman et al. (2017); Tang et al. (2024); Ullah et al. (2020) more commonly used words such as 'the', 'is', 'and', 'or' are removed to reduce the impact of less important words. Stemming Alawneh et al. (2024); Alfreihat et al. (2024); Ayvaz and Shiha (2017); Khan et al. (2025); Kralj Novak et al. (2015); Tang et al. (2024) reduces words to their stem by removing the suffix, and lemmatization Jona et al. (2022); Khan et al. (2025) uses grammar rules to reduce the words to their base form. Both the processes reduces the redundancy. Spell correction is an obvious process for cleaning the text. Slangs are removed and informal spellings are handled using the process of normalization.

Part of speech tagging, Hogenboom et al. (2013); Lou et al. (2020) an advanced preprocessing technique, tags the text into grammatical roles such as nouns, pronouns, adjectives and so on. Named Entity recognition, often known as NER identifies the real world entities.

Emojis are either removed from the text or handled in a special manner by converting them to a descriptive token, often sentimental. Different ways of handling the emojis are elaborated in the next section.

4.1 Special Processing Techniques of Emojis

Emojis play a vital role in digital communication, significantly influencing the interpretation of sentiments. Re-search reveals that a single emoji can be associated with multiple functionalities, i.e. it may emphasize existing sentiment,

Table 2: *Preprocessing Techniques*

Preprocessing step	Reference
Stopword removal	Alfreihat et al. (2024); Ayvaz and Shiha (2017); Jag adishwari et al. (2021); Jona et al. (2022); Khan et al. (2025); Kralj Novak et al. (2015); Lou et al. (2020); Rahman et al. (2017); Tang et al. (2024); Ullah et al. (2020)
Stemming	Alawneh et al. (2024); Alfreihat et al. (2024); Ayvaz and Shiha (2017); Khan et al. (2025); Kralj Novak et al. (2015); Tang et al. (2024)
Lemmatization	Jona et al. (2022); Khan et al. (2025)
Part-of-speech (POS)	Hogenboom et al. (2013); Lou et al. (2020)
Tokenization	Alawneh et al. (2024); Alfreihat et al. (2024); Ayvaz and Shiha (2017); Hogenboom et al. (2013); Jagadishwari et al. (2021); Karthikeya et al. (2023); Khan et al. (2025); Kralj Novak et al. (2015); Lou et al. (2020); Mona M. Abd Elsalam (2022); Rahman et al. (2017); Singh et al. (2024); Srisankar and Lochanambal (2021); Tang et al. (2024); Ullah et al. (2020); Vashisht and Jailia (2020); Yin et al. (2023); Yoo and Rayz (2021); Zhao et al. (2024)
Normalization	Alfreihat et al. (2024)

introduce new sentiment, mitigate harshness, reverse the polarity of a message, trigger positive or negative reactions, or remain neutral. For instance, the same emoji might intensify positivity in one context (e.g., -Perfect! :)-) but can convey negativity in some other context (e.g., -Not great :)-). Apart from modifying the meaning of the sentences, a single emoji can often convey the meaning of the entire sentence. These techniques can be further categorized into the following distinct approaches. (ref Fig. 1)

Emoji Sentiment Lexicons are predefined dictionaries that often assign fixed sentiment scores (e.g., -1 to +1) or polarity labels (positive/ negative/ neutral) to emojis based on human annotations. For instance, a smiley emoji can be assigned a high positive score of +0.97 while a strong negative score of -0.97 can be assigned to a sad face emoji. The emojis are

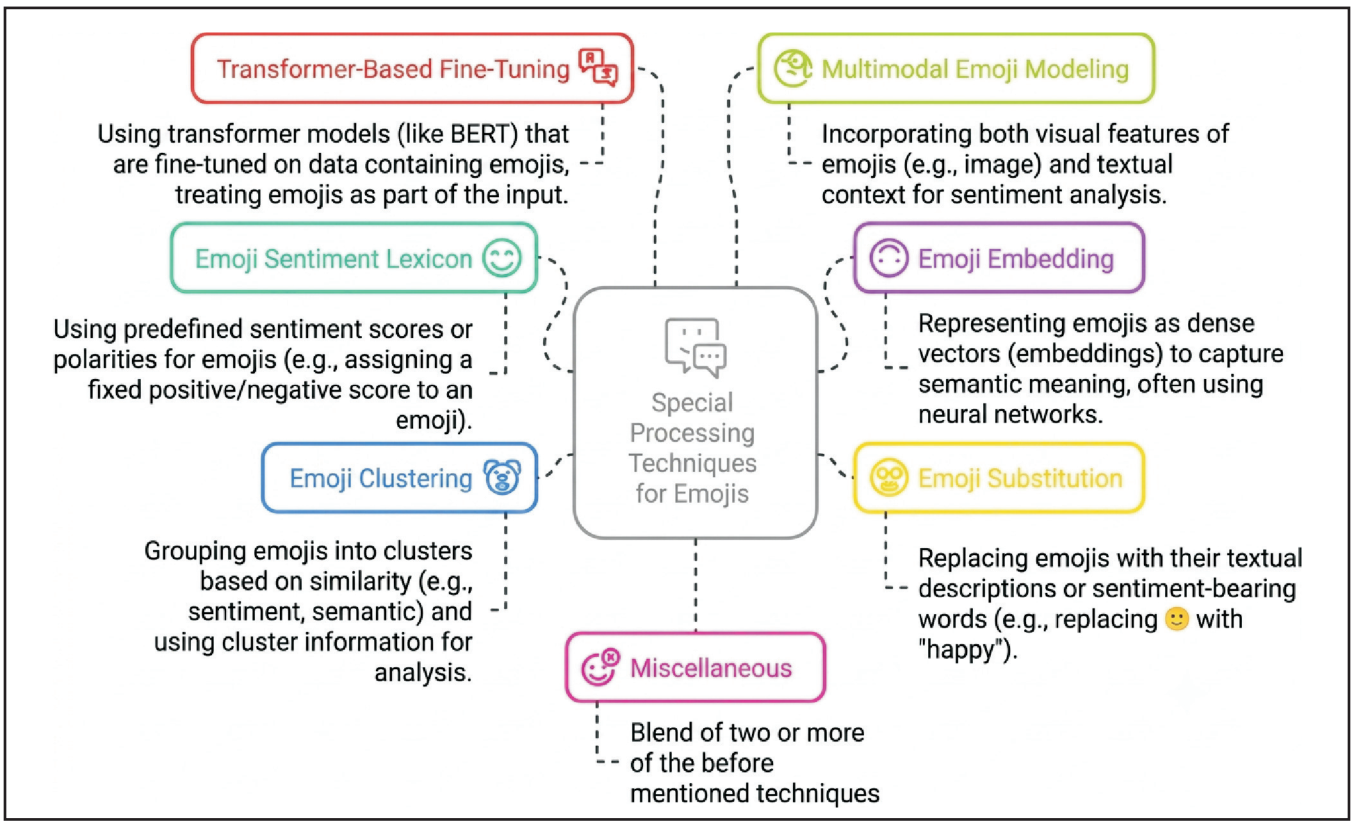

Fig. 1: *Techniques used for emoji processing.*

then replaced by their assigned scores or labels, thereby allowing incorporation of these emojis with the traditional classifiers. The studies Alfreihat et al. (2024); Ayvaz and Shiha (2017); Godard and Holtzman (2022); Hakami et al. (2022); Hogenboom et al. (2013); Holthoff (2020); Karthikeya et al. (2023); Khan et al. (2025); Santos et al. (2023); Singh et al. (2024); Srisankar and Lochanambal (2021); Toet et al. (2018); Vashisht and Jailia (2020); Yoo and Rayz (2021) use this technique for handling the emoji rich text.

Emoji Embedding techniques convert emojis into dense vector representations by using algorithms such as Word2Vec, FastText, or contextual models (BERT). These embeddings capture semantic similarities, grouping related emojis (e.g., heart emoji and face with hearts in eyes emoji cluster together due to shared emotional connotations). This method also encodes multidimensionality (e.g., fire emoji simultaneously conveys threat, hype, or cultural relevance). The studies Ayvaz and Shiha (2017); et al. (2020); Karthikeya et al. (2023); Mona M. Abd Elsalam (2022); Singh et al. (2024); Srisankar and Lochanambal (2021); Wang and Castanon (2019); Yoo and Rayz (2021); Yuan et al. (2022); Zhao et al. (2024) demonstrated the use of this technique for processing of emojis.

Emoji Substitution can convert emojis into machine-processable tokens by replacing emojis with text equivalents. For instance, explicit labels (e.g., ″[happy face]″ are used for happy face emoji) or contextual keywords (e.g., ″cele-bration item″ for its corresponding emoji). This substitution allows the application of traditional NLP techniques for processing of this emojis. Alawneh et al. (2024); Jagadishwari et al. (2021); Karthikeya et al. (2023); Khan et al. (2025); Rahman et al. (2017); Yoo and Rayz (2021)

Emoji Clustering groups semantically or sentimentally similar emojis using unsupervised algorithms like K-means or hierarchical clustering, typically applied to pre-trained emoji embeddings. For example, clustering might group laughing emojis into a ″joy/humor″ cluster, while all anger related emojis form an ″anger″ cluster. These clusters serve as categorical features in sentiment models, reducing dimensionality and generalizing across visually diverse but functionally equivalent emojis. This technique simplifies analysis by treating emoji variants (e.g., skin tones) as a single semantic unit. Alfreihat et al. (2024); Hakami et al. (2022); Wang and Castanon (2019); Yuan et al. (2022)

Transformer-Based Fine-Tuning for Emojis This approach fine-tunes pretrained transformer models (e.g., BERT, RoBERTa) on domain specific, emoji rich datasets, treating emojis as first-class tokens during training. The model learns context-dependent representations by analyzing emojis alongside surrounding text e.g., distinguishing positive sentiment in ″Aced the exam! party popper emoji″ from negative connotations in ″Bankrupt party popper emoji.″ Unlike static lexicons or embeddings, transformers capture sarcasm, irony, and cultural nuances by weighing emojis within full-sentence contexts. et al. (2020); Khan et al. (2025); Singh et al. (2024); Tang et al. (2024); Zhao et al. (2024)

Multimodal Emoji Modeling jointly analyzes the visual properties of emojis (e.g., color, shape) and their textual context using architectures like ViLBERT or CLIP. These models fuse image features (extracted via CNNs) with text embeddings to interpret emojis holistically e.g., recognizing that ″leaf emoji″ might signify nature (green leaves) or marijuana (cultural symbolism) based on combined visual-textual signals. This is critical for emojis where meaning derives from imagery (e.g., spark emoji as ″explosion″ or ″impact″). Applications include sentiment analysis in platforms where emojis substitute words (e.g., messaging apps). Jagadishwari et al. (2021); Lou et al. (2020); Zhao et al. (2024)

4.2 Advanced Social Media Preprocessing

Social media text is often unstructured and informal. This makes it difficult to interpret the sentiment correctly. This issue can be addressed by using the following preprocessing techniques.

- **Unicode Normalization and Standardization:** Since emoticons and emojis can be represented differently across the different social media platforms, Unicode Normalization is often used to convert them into astandard format Kralj Novak et al. (2015). In some studies, emojis are converted into their standard form before tokenization. This leads to the preservation of their emotional menaing Kralj Novak et al. (2015).
- **Handling Code-Mixed Text:** Maany rich emojis messages contain code-mixed text, where two or more langauges are used within the same message. This is commonly used in the *Sentimix* dataset; which is specifically designed for multi-lingual texts Yuan et al. (2022). Specialized embedding techniques are used for processing such datasets.
- **Hashtag and Keyword Parsing:** Social media data is often collected using hashtags and keywords Singh et al. (2024). During preprocessing, these hashtags are separated from the main text and handled individually Singh et al. (2024). Since, hashtags can contain multiple words (e.g., #NewYearsEve), they are split into individual words Vashisht and Jailia (2020).
- **Normalization of Informal Language:** Digital text often contains informal words, slangs and abbreviations. Text Normalization is used for converting these into their standard meanings. This in turn preserves their sentiments Zhao et al. (2024); Toet et al. (2018); Kralj Novak et al. (2015).

Miscellaneous In addition to the methods described above, there are other approaches to preprocessing social media texts that incorporate some form of emoji recognition and analysis. These hybrid approaches leverage complementary strengths to overcome limitations of single-method implementations. The study Santos et al. (2023) proposes an EmojiMapper function for its preprocessing filters to map emojis to textual meanings, after which a lexicon was used to determine sentiments of the messages. The study Ullah et al. (2020) took a phased methodology: first using 732 hand-classified emoticons (positive/negative/neutral) as lexicon features alongside text representations, then removing them entirely to measure performance impact. This strategy prioritizes interpretability while mitigating context-blindness through comparative analysis. The study Jona et al. (2022) uses the blend of the embeddings and modeling synergy leveraging deep learning to capture contextual dynamics through the use of DeepMoji SE. Emojis were retained as contextual signals during transformer-based fine-tuning on GitHub data, generating emotion-rich embeddings that fed into their SEntiMoji classifier. In later work, they introduced EmoGraph2vec, creating graph-based emoji embeddings fused with text via hybrid attention mechanisms before classification through TextCNN architectures. These approaches treat emojis as semantic anchors, using their distributional properties to bootstrap contextual understanding rather than relying on fixed dictionaries.

5. Feature Extraction Techniques

Text having emojis is converted into features that can be used further for classification in various machine learn-ing models (refer Table 3. Approaches like Bag of words Ayvaz and Shiha (2017); Jagadishwari et al. (2021); Karthikeya et al. (2023); Mona M. Abd Elsalam (2022); Srisankar and Lochanambal (2021); Wang and Castanon (2019), TF-IDF Ayvaz and Shiha (2017); et al. (2020); Jona et al. (2022); Karthikeya et al. (2023); Khan et al. (2025); Rahman et al. (2017); Santos et al. (2023); Singh et al. (2024); Toet et al. (2018); Wang and Castanon (2019), ngram Toet et al. (2018); Vashisht and Jailia (2020) have been used to extract tokens from the text. However, several approaches like emoji sentiment ranking Godard and Holtzman (2022); Hakami et al. (2022); Hogenboom et al. (2013); Kralj Novak et al. (2015); Toet et al. (2018); Ullah et al. (2020) and AFINN are used for extraction of tokens from emojis. A lexicon based approach, emoji sentiment ranking provides sentimet scores for indi-vidual emojis. A large dataset of 1.6 million tweets having emojis was used, and each emoji in the dataset was assigned a polarity value depending on the occurrence of the same in tweets. Sentiment score was calculated using the formula Sentiment Score = Positive Probability - Negative Probability. Another lexicon based approach, AFINN asssigns predefined sentiment scores to English words, ranging from -5 (very negative) to +5 (very positive).

Word embeddings like Word2Vec Alfreihat et al. (2024); Khan et al. (2025); Yoo and Rayz (2021); Yuan et al. (2022), GloVe Karthikeya et al. (2023); Wang and Castanon (2019), Emoji2Vec Godard and Holtzman (2022); Holthoff (2020), deep-moji Godard and Holtzman (2022) are deep learning-based approaches that capture semantic relationships. Several Emoji-aware embeddings, Emoji2Vec Godard and Holtzman (2022); Holthoff (2020), FastText have been used by researchers that combine pre-trained word embeddings with emoji embeddings.

6. Classification Techniques

Several machine learning (ML) and deep learning techniques are used for sentiment classification using Emojis, mentioned in Table 3. Some of the traditional machine learning methods include multinomial Naive Bayes Alfreihatet al. (2024); et al. (2020); Jagadishwari et al. (2021); Mona M. Abd Elsalam (2022); Santos et al. (2023); Toet et al. (2018); Ullah et al. (2020); Vashisht and Jailia (2020), SVM Alfreihat et al. (2024); Ayvaz and Shiha (2017); et al. (2020); Hakami et al. (2022); Jagadishwari et al. (2021); Karthikeya et al. (2023); Lou et al. (2020); Mona M. Abd Elsalam (2022); Santos et al. (2023); Toet et al. (2018); Ullah et al. (2020); Vashisht and Jailia (2020), Random Forest et al. (2020); Mona M. Abd Elsalam (2022); Ullah et al. (2020), K-Nearest Neighbors Hakami et al. (2022); Mona M. Abd Elsalam (2022); Ullah et al. (2020), and Logistic Regression Alfreihat et al. (2024); Jagadishwari et al. (2021); Mona M. Abd Elsalam (2022). Most of these methods use features like word frequency, TF-IDF, emojis sentiment score, sentiment lexicons, part of speech tags.

Deep learning techniques automatically extract features from text data that include emojis. Some of the commonly used techniques are Convolutional Neural Networks (CNNs) Alawneh et al. (2024); Hakami et al. (2022); Karthikeya et al. (2023); Lou et al. (2020), Recurrent Neural Networks (RNNs), LSTM (Long Short-Term Memory) Alawneh et al. (2024); Hakami et al. (2022); Karthikeya et al. (2023); Lou et al. (2020); Tang et al. (2024), GRU (Gated Recurrent Units) Alfreihat et al. (2024), BiLSTM + Attention Yuan et al. (2022), Emoji2Vec / Word2Vec + LSTM. While CNNs works on local n-gram features, RNNs work on sequential information in text. LSTM, a type of RNN, remembers long-term dependencies and handles varying input lengths, making it suitable for texts having emojis. GRU, a faster implementation of LSTM, is effective in handling sequential data like text. BiLSTM and Attention mechanisms focus on the relevance part of the text.

Further, advancements in transformer-based models have enhanced the capabilities to understand the context of text. Bidirectional Encoder Representations from Transformers (BERT) model Khan et al. (2025) handles the context-dependent emoji usage in the text. RoBERTa, DistilBERT, XLNet are the lighter varients of BERT. Emoji-BERT is specially pre-trained on emoji dataset.

6.1 Evaluation Techniques

Most of the researchers have used accuracy as the evaluation metric for traditional datasets and , precision, recall and F1 score for imbalanced datasets. Ayvaz and Shiha (2017); et al. (2020); Hogenboom et al. (2013); Holthoff (2020); Jagadishwari et al. (2021); Karthikeya et al. (2023); Khan et al. (2025); Lou et al. (2020); Mona M. Abd Elsalam (2022); Singh et al. (2024); Tang et al. (2024); Ullah et al. (2020); Vashisht and Jailia (2020); Yin et al. (2023); Yuan et al. (2022) However, other evaluation metrics are also used like Jaccard index for Emoji ranking Singh et al. (2024) and overlap analysis, hamming loss for Multi-label emoji tagging Singh et al. (2024), Top k-accuracy for Emoji recommendation/prediction.

Table 3: *Feature Extraction and ML/DL Techniques*

Ref	Task / Language	Dataset Size	Feature Extraction	ML/DL Methods	Best Performance
[1]	Arabic SA	56,000	Word Embedding (Emoji Encoding)	CNN-LSTM	Acc (91.85%), F1(0.92)
[2]	Arabic SA	58,000	Word2Vec, Emoji Feature Vectors	Bi-GRU, SVM, NB, LR	Acc (89%), F1(0.89)
[3]	English SA	175,181	BoW, TF-IDF, Word Embeddings	SVM, Neural Networks	Acc (81%), F1(0.80)
[4]	English SA	NES	TF-IDF, Word Embedding	SVM, Naive Bayes, RF	Acc (82.10%), F1(0.82)
[5]	English Emotion	3,000,000	NRC Lexicon, Emoji2Vec, Deep-moji	Deep CNN, MobileNet	Acc (79.1%), F1(0.79)
[6]	Arabic SA	144,196	ASAD, CAMEL, MAZAJAK, Lexicon	SVM, KNN, CNN, LSTM	Acc (85.3%), F1 (0.85)
[7]	Dutch, English SA	2,080	Emoji Sentiment Lexicon	Lexicon-based	Acc (74.2%), F1(0.75)
[8]	English SA	NES	Emoji2Vec	Unsupervised Lexicon SA	Acc (81.43%), F1 (0.81)
[9]	English SA	3,000,000	BoW	Multinomial NB, SVM, Regr.	Acc (86.4%), F1 (0.86)
[10]	English Prediction	NES	TF-IDF	CNN, Bi-LSTM, Softmax, RNN	Acc (82%), F1 (0.82)
[11]	English SA	NES	BoW, TF-IDF, GloVe	SVM, CNN, LSTM	Acc (87%), F1 (0.87)
[12]	English SA (Airlines)	14,640	TF-IDF, Word2 Vec, BERT	BERT, SVM, RF, NB, LR	Acc (88%), F1 (0.88)
[13]	English SA	1,600,000	Lexicon Matching, Rule Weighting	Custom Rule-based	Acc (76.8%), F1 (0.76)
[14]	Chinese/English SA	10,042	Word Embedding	CNN, LSTM, SVM	Acc (96.91%), F1 (0.92)
[15]	English SA (COVID)	10,000	BoW	SVM, RF, GB, KNN, GNB	Acc (79.6%), F1 (0.79)
[16]	Multiling. Emotion	460 KW	TF-IDF	KNN, Rule-based, PMI	Acc (88%), F1 (0.82)
[17]	Portuguese, Eng SA	84,115	TF-IDF, Word Embeddings	SVM, Naive Bayes	Acc (95%), F1 (0.91)
[18]	Code-mixed SA	20,000	Char/Elmo Embeddings, TF-IDF	Multi-task, Transformer	Acc (86%), F1 (0.84)
[19]	13 Languages SA	1,600,000	BoW	CNN, NB, SVM, Max Ent	Acc (80%), F1 (0.82)
[20]	English SA	Multi-DS	Word Embeddings, Topic Model	LSTM, Bi-LSTM	Acc (91%), F1 (0.93)
[21]	Food Emotion	60 Images	TF-IDF, N-grams, Emoji Lexicon	NB, SVM, Decision Tree	Acc (73.4%), F1 (0.76)
[22]	English SA (Airline)	14,460	Emoji/NRC Lexicons, NileUlex	SVM, NB, KNN, RF	Acc (87%), F1 (0.86)
[23]	English SA	NES	N-grams (Unigrams, Bi-grams)	SVM, Max Ent, NB	Acc (83.6%), F1 (0.81)
[24]	English Prediction	13,000,000	BoW, TF-IDF, GloVe	CNN, MNB, SVM, RNN	Acc (84%), F1 (0.82)
[25]	English SA	184,485,919	Distant Supervision	CNN, Transfer Learning	Acc (78%), F1 (0.8)
[26]	English SA	1,800,000	Word2Vec (CBOW/Skip-Gram)	CNN	Acc (72%), F1 (0.82)
[27]	English SA	NES	Word2Vec	Bi-LSTM, Hybrid Attention	Acc (89%), F1 (0.9)
[28]	Chinese SA	6,100,000	Manual Interpretation (EmojiGrid)	TaneNet with Attention	Acc (81.43%), F1 (82)

PMI-based classification refers to a sentiment or text classification approach that uses Pointwise Mutual Information (PMI) to measure the association between words (or phrases) and categories like positive, negative, or specific topics. PMI is often used to calculate how strongly a word is associated with positive or negative sentiment.

Emoji vectors and Word2Vec can overlap, but only if emojis are included in the training corpus and treated astokens. Otherwise, they are generated separately.

7. Quantitative Analysis of Research Trends

Based on a systematic review of the literature, several broad claims regarding the current state of sentiment analysis in emoji-rich text are supported by specific counts and trends identified among the 28 primary research papers analyzed.

- Dominance of Social Media and Twitter: Majority of the emoji-rich datasets are often curated from the social media. Among them, Twitter is the most commonly used platform; utilised by 12 studies Mona M. Abd Elsalam (2022); Ayvaz and Shiha (2017); Hakami et al. (2022); Wang and Castanon (2019); Jagadishwari et al. (2021); Lou et al. (2020); Tang et al. (2024); Vashisht and Jailia (2020); Alfreihat et al. (2024); Zhao et al. (2024); Rahman et al. (2017); Hogenboom et al. (2013). A smaller subset of the studies utilizes data from Facebook, Gaming forums and Sina Weibo Godard and Holtzman (2022); et al. (2020); Zhao et al. (2024); Alawneh et al. (2024).
- Variability in Dataset Scale and Acquisition: There is a considerable variation both in the size of the datasets and methods used for data collection. Size of the dataset ranges from FRIDa database containing only 60 images to a large scale dataset comprising of 184 million tweets Kralj Novak et al. (2015); Ullah et al. (2020). Among the reviewed studies, 17 collected data using web-scraping, while 11 used pre-compiled datasets Singh et al. (2024).
- Linguistic Diversity and Trends: English remains the most widely used language in the existing studies. However, gradually research is expanding and becoming more diversified. Holthoff (2020) uses 13 languages. Chinese is also widely studied through the microblogs collection from San Weibo et al. (2020); Alawneh et al. (2024). Similarly, Arabic has also gained significant attention and used in Yoo and Rayz (2021); Santos et al. (2023); Yin et al. (2023).
- Standardization of Preprocessing pipelines: Tokenization is the most widely used technique reported in 19 studies. Stopword Removal is the next commonly utilized technique used in 10 studies Ullah et al. (2020).
- Evaluation Metric: Accuracy is the standard metric often used for balanced datasets. Other metrics like Recall, F1-score and Precision are also used in 15 of the existing studies.
- Prevalance of Machine Learning Architectures: Support Vector Machine is the most commonly used traditional classifier used in 12 of the existing studies. Multinomial Naive Bayes is the next commonly used classifer reported in 8 studies, More advanced architectures like CNNs and LSTMs are also utilized.

8. Qualitative Analysis of Research Trends

- Performance vs Computational Trade-offs: Transformer-based models often achieve more accuracy as compared to the traditional classifiers. But they require more computational resources and training time.
- Scalability and Dataset Biases: Traditional ML models are trained using static and handcrafted features; which can easily be mimicked by the adversarial entities. Also, using Generative AI introduces latency; enforcing a limit on the scalability.
- Precision-Recall Tradeoffs: Using complex models like LLMs often lead to the increase in recall but causes precision to drop. This is mainly because the model becomes more hyper-sensitive to multilingual cues.
- Linguistic Evaluation: Stemming often leads to sparsity in complex languages like Arabic. This gap is addressed by translation of emojis into textual descriptions. This is turn allows the deep learning layers to make use of their pretraining vocabulary.

9. Challenges in Handling of Emojis

Emojis have high expressive power, yet they are treated as noise by many. Most researchers ignore emoji handling while dealing with the text due to the following challenges:

- Language and Regional Variability: Different regions, cultures have different ways of expressing the emotions. Emojis also fall in that category. Emojis do not always carry fixed meanings. Same emoji may be expressing different emotions in different regions and languages Khan et al. (2025); Mona M. Abd Elsalam (2022) . This leads to ambiguity and bias in the models. Most of the research work in this area is related to English language Santos et al. (2023). The same tools or techniques may not give accurate results for other languages. There is a strong need to have language specific emoji interpretations and cultural variability Yin et al. (2023).
- Context dependence: The emotion of an emoji is dependent on the context Khan et al. (2025); Mona M. Abd Elsalam (2022) . An emoji can have multiple meanings, known as Polysemy, e.g. the "Red Heart" emoji can convey anger, hurt, rage, or madness and a keyword can be represented by multiple emojis, known as Synonymy Singh et al. (2024). Such emojis are difficult to interpret.
- Lack of Standard Representation: Often, emojis lack a proper way of vectorization. Most of the existing dictionaries ignore emojis. There is no universal dictionary, hence it is difficult to assign consistent sentiment scores. There is a need to have a robust emojis dictionary and proper mechanism for conversion of emojis to numeric forms, which can be further used in machine learning algorithms Mona M. Abd Elsalam (2022); Khan et al. (2025) . Further, use of joint emojis have increased with time but mechanism to understand the meaning of

joint emojis is missing in practice Santos et al. (2023). Emoticons and emojis are sometimes used sarcastically and lead to the creation of noisy labels in distant supervision Yin et al. (2023). There is no standard emojis lexicons e.g. Santos et al. (2023) uses one of more of a 100 emojis defined in their model.

- Short Text Issues: Recently, people have started using short texts having more emojis and emoticons to communicate. Such texts are casual and noisy and lack semantic expressions, leading to ambiguous interpretations Tang et al. (2024). Conversion of such texts leads to sparse features, making it difficult for NLP techniques to interpret the emotions or expressions from the same. Further, being small texts, emojis used in these texts should have more weightage than the emojis used in longer documents Khan et al. (2025).
- Model Integration of Text and Emoji: Text and emoji have different characteristics and need to be handled differently. However, current models combine both of them in the same model, making it ineffective for sentiment classification. Often, current models either ignore or misinterpret emojis when they are dealt with text Mona M. Abd Elsalam (2022).
- Generative AI Approach: Some of the pretrained language models (PLM) are not trained on emojis Tang et al. (2024) and the models trained on emojis are computationally expensive. They perform on emoji-aware input but training and fine tuning such models require heavy resources Khan et al. (2025).

10. Research Gaps and Synthesis

Despite significant advances in this research field, several research gaps still remain unresolved. Existing models can not recognize the change in the semantics of the emojis based on its intent or surrounding text. This issue is further worsened by the linguistic and the cultural bias present in the existing studies. Furthermore, non-availability of well labeled, large and multilingual datasets is another issue. Also, majority of the existing studies treat emojis as text and ignore their visual appearance. They also pay little attention to repeated or grouped emojis. In addition, different analysis methodologies often produce varied results; making it difficult to compare results. This also highlights the need for a standard techniques. Future studies shall involve multimodal approaches; thereby integrating user behavior, text and emojis to have a better understanding of digital communication.

10.1 RQ1: Characteristics of Existing Datasets

According to the analysis drawn from the literature, social media emerges as the primary source of emoji centric data. **Twitter** stands out to be the most promising one with the highest number of citations. This is mainly due to its open access and robust APIs Yoo and Rayz (2021).

- **Size and Scope:** The size of the datasets are extremely inconsistent in size, spanning from small focused collections of **60 images** to massive datasets such as the **3.8 TB Twitter Stream Grab** containing millions of entries Santos et al. (2023); Mona M. Abd Elsalam (2022).
- **Acquisition:** Researchers have typically accessed these datasets either by **direct downloads** (11 studies) or by **web-scraping** (17 studies). Scraping is often preferred for its capability to capture the niche data with respect to specific events such as airline customer service or Covid-19 pandemic Ayvaz and Shiha (2017); Hakami et al. (2022).
- **Linguistic Diversity:** Although English is the baseline, research is expanding to Arabic, Chinese (Sina Weibo) and Portuguese, with a study covering **13 distinct languages** Yin et al. (2023).

10.2 RQ2: Preprocessing, Feature Extraction, and Machine Learning Techniques

The processing pipeline for emoji-integrated text adheres to a structured sequence of data cleaning and vectorization.

- **Preprocessing: Tokenization** emerges as a most foundational processing step, reported in 19 studies, followed by **stopword removal** in 10 studies Srisankar and Lochanambal (2021). Additional techniques include stemming, lemmatization, and normalization Srisankar and Lochanambal (2021); Godard and Holtzman (2022).
- **Feature Extraction:** Conventional studies often utilize techniques like **Bag-of-Words (BoW)** and **TF-IDF** Wang and Castanon (2019). Conversely, more advanced studies exhibit the usage of **word and emoji embeddings** including Word2Vec, GloVe, and specialized vectors like **Emoji2Vec** Jagadishwari et al. (2021).
- **ML/DL Models:** Classic machine learning algorithms like **SVM** and **Naive Bayes** remain popular due to their consistency Jagadishwari et al. (2021). However, **Deep Learning** models (CNNs, LSTMs, and GRUs) are preferred for modeling temporal dependencies, while **Transformers techniques** like BERT, Emoji-BERT are utilize to better capture the contextual nuances in the text Lou et al. (2020); Singh et al. (2024).

10.3 RQ3: Specialized Methods for Handling Emojis

Four primary specialized strategies are employed to integrate emojis into sentiment analysis Tang et al. (2024):

1. **Preprocessing:** Emojis are either substituted with descriptive text tokens or converted into standardized Unicode equivalents Tang et al. (2024); et al. (2020).

2. **Lexicon-Based Mapping:** Assigning fixed polarity scores from dictionaries like the **Emoji Sentiment Lexicon (ESL)** or **EmojiNet** Vashisht and Jailia (2020); Karthikeya et al. (2023).
3. **Embedding Representations:** Capturing semantic relationships through dense vector spaces (static or contextual) that allow the emoji's meaning to adapt to its sentence context Jona et al. (2022).
4. **Modeling Architectures:** Utilizing specialized neural designs like **BiLSTMs with attention** or **TaneNet** to dynamically weight emoji influence Yuan et al. (2022); Holthoff (2020).

10.4 RQ4: Key Challenges and Research Gaps

Despite recent progress, research on emoji analysis continues to encounter several formidable challenges:

- **Challenges:** Emojis are highly **context-dependent (polysemy)** and subject to **regional/cultural variability**, leading to ambiguity and bias Khan et al. (2025). There is also a distinct **lack of standard representation** and universal dictionaries Khan et al. (2025).
- **Research Gaps:** Future work is needed in **context-aware interpretation** beyond static mapping Ullah et al. (2020). Most studies only explore **three-way polarity** (positive, negative, neutral), leaving granular emotions like irony and sarcasm underexplored Ullah et al. (2020). Furthermore, there is a lack of **multimodal approaches** that integrate emojis with images, video, and audio Ullah et al. (2020); Alfreihat et al. (2024).

11. Conclusion and Future Work

In this paper, authors have discussed about various methods of handling text with emojis and emoticons. The study has been divided into sections related to datasets available, feature engineering methods, machine leaning models. Various Challenges and research gaps in this area of research are discussed. The area has lots of scope in future. A roadmap for researchers in the area of emoji handling will be to create a strong and elaborate dictionary for emojis, which can handle polysemy and Synonymy. Context dependence of emoji interpretation and linguistic variablity should be handled in a robust manner. Dependence of culture and language needs to be handled. Use of emojis depend on the age, gender and location, Such factors should be considered while converting emojis to numeric forms. The model needs to be developed to take care of the aspect of emojis evolution with time. Rise of new emojis and changing usage pattern should be handled by the models based on continual learning. Specific eventse.g. elections, pandemic, give rise to certain sentiments. Such kind of longitudinal event analysis needs to be done. With the advancements in generative AI approaches, emoji-aware pretrained language models should be developed. Based on the nature of the data, more weightage may be give to the importance of emojis and develop the machine learning models to handle such implementations. Integration of text, emojis, images, audio, video, and other media is the need of the hour.

Conflict of Interest Statement

The authors declare no conflict of interest.

Ethical Declarations

The authors declare that this research does not involve human participants, animals, or any procedures requiring ethical approval. All data used in this study were obtained and processed in accordance with applicable ethical andlegal guidelines.

References

1. Alawneh, H., Hasasneh, A., and Maree, M. (2024). On the utilization of emoji encoding and data preprocessing with a combined cnn-lstm framework for arabic sentiment analysis. *Modelling*, 5:1469–1489.
2. Alfreihat, M., Almousa, O. S., Tashtoush, Y., Alsobeh, A., Mansour, K., and Migdady, H. (2024). Emo-sl framework: Emoji sentiment lexicon using text-based features and machine learning for sentiment analysis. *IEEE Access*, 12:81793–81806.
3. Ayvaz, S. and Shiha, M. (2017). The effects of emoji in sentiment analysis. *International Journal of Computer and Electrical Engineering*, 9:360–369.
4. Et al., A. S. (2020). Alternative method sentiment analysis using emojis and emoticons. *Procedia Computer Science*, 178:182–193.
5. Godard, R. and Holtzman, S. (2022). The multidimensional lexicon of emojis: A new tool to assess the emotional content of emojis. *Frontiers in Psychology*, 13:921388. https://www.frontiersin.org/articles/10.3389/fpsyg.2022.921388/pdf.
6. Hakami, S. A. A., Hendley, R., and Smith, P. (2022). Emoji sentiment roles for sentiment analysis: A case study in Arabic texts. In Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 346–355, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
7. Hogenboom, A., Bal, D., Frasincar, F., Kaymak, U., and de Jong, F. (2013). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC)*, pages 703–710. ACM.

8. Holthoff, L. (2020). The emoji sentiment lexicon: Analysing consumer emotions in social media communication. In *Proceedings of the European Marketing Academy*, volume 49, page 63029.
9. Jagadishwari, V., Indulekha, A., Raghu, K., and Harshini, P. (2021). Sentiment analysis of social media text-emoticon post with machine learning models contribution title. *Journal of Physics: Conference Series*, 2070(1):012079.
10. Jona, J., Jothi, C., Paul, S. D., Gunasekaran, S., and Sridhar (2022). Sentimental analysis using emojis prediction. *International Journal of Research Publication and Reviews*, 3(6):1245–1249.
11. Karthikeya, K., Yaswanthi, K., Sai, E. J., Pradeepini, G., Kiran, Y. N. V. N. A. S., and Bulla, S. (2023). Sentiment analysis of tweets using logistic regression and neural networks with emojis and emoticons. *Research Square*.Khan, A., Majumdar, D., and Mondal, B. (2025). Sentiment analysis of emoji fused reviews using machine learning and bert. *Scientific Reports*, 15:7538.
12. Kralj Novak, P., Smailovic´, J., Sluban, B., and Mozeticˇ, I. (2015). Sentiment of emojis. *PLOS ONE*, 10(12):1–22.
13. Lou, Y., Zhang, Y., Li, F., Qian, T., and Ji, D. (2020). Emoji-based sentiment analysis using attention networks. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5):1–13.
14. Mona M. Abd Elsalam, Ahmed M. Gadallah, H. A. H. (2022). Sentiment analysis of text incorporating emojis: A machine learning approach. *International Journal of Computer Applications*, 184(20):10–18.
15. Rahman, R., Islam, T., and Ahmed, M. H. (2017). Detecting emotion from text and emoticon. *London Journal of Research in Computer Science and Technology*, 17(3):1–13.
16. Santos, V. M. d. S., Freitag, R., Tejada, J., Floreˆncio, A. S., Souza, P. P. O. B., and Gois, T. S. (2023). Sentiment analysis with emojis: a model for Brazilian. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 613–616, Vienna, Austria. NOVA CLUNL, Portugal.
17. Singh, G., Ghosh, S., Firdaus, M., Ekbal, A., and Bhattacharyya, P. (2024). Predicting multi-label emojis, emotions, and sentiments in code-mixed texts using an emojifying sentiments framework. *Scientific Reports*, 14.
18. Srisankar, M. and Lochanambal, K. (2021). Emoji's impact on sentiment analysis - a survey. *International Journal of Innovative Research in Technology*, 8(7):366–371.
19. Tang, H., Tang, W., Zhu, D., Wang, S., Wang, Y., and Wang, L. (2024). EMFSA: Emoji-based multifeature fusion sentiment analysis. *PLOS ONE*, 19(9):e0310715.
20. Toet, A., Kaneko, D., Ushiama, S., Hoving, S., de Kruijf, I., Brouwer, A.-M., Kallen, V., and van Erp, J. B. F. (2018). Emojigrid: A 2d pictorial scale for the assessment of food-elicited emotions. *Frontiers in Psychology*, 9:2396.
21. Ullah, M. A., Marium, S. M., Begum, S. A., and Dipa, N. S. (2020). An algorithm and method for sentiment analysis using the text and emoticon. *ICT Express*, 6:357–360.
22. Vashisht, G. and Jailia, M. (2020). Emoticons emojis based sentiment analysis: The last two decades! *International Journal of Scientific Technology Research*, 9(03):6398–6405.
23. Wang, Y. and Castanon, J. (2019). Emojify: Emoji prediction from sentence. CS229: Machine Learning Projects, Stanford University.
24. Yin, W., Alkhalifa, R., and Zubiaga, A. (2023). The emojification of sentiment on social media: Collection and analysis of a longitudinal twitter sentiment dataset.
25. Yoo, B. and Rayz, J. (2021). Understanding emojis for sentiment analysis. *The International FLAIRS Conference Proceedings*, 34.
26. Yuan, X., Hu, J., Zhang, X., and Lv, H. (2022). Pay attention to emoji: Feature fusion network with emograph2vec model for sentiment analysis.
27. Zhao, Q., Wu, P., Lian, J., An, D., and Li, M. (2024). Tanenet: Two-level attention network based on emojis for sentiment analysis. *IEEE Access*, 12:86106–86119.

❖ ❖ ❖