International Journal of Engineering and Techniques - Volume 12 Issue 2, March-April - 2026

MULTI-TASK EMOJI LEARNING

Paramita Dey[1],Soumya Dey[2]

[1] Department of Information Technology, Government College of Engineering and Ceramic Technology

[2] Department of Information Technology, Government College of Engineering & Ceramic Technology

Abstract

Twitter is a social networking platform where users can create, view, and interact with short messages known as "tweets." This large volume of user-generated content can be utilized by organizations, such as businesses, to understand customer opinions and attitudes. This research paper focuses on the use of emoticons in social media and the emotions they express. The aim of the study is to present a framework for analyzing emotional responses in real-world Twitter data. The proposed approach is based on supervised machine learning techniques and uses data collected through the "TWEEPY" crawler for experimental analysis. The gathered dataset is preprocessed and refined before being applied to different supervised models. Finally, each tweet is classified according to the emotional sentiment of the user, categorized as positive, negative, or neutral.

Keywords

Twitter, machine learning, sentiment analysis, opinion mining, Random Forest (RF), Decision Tree (DT), Bag-of-Words (BoW) approach, TF–IDF (Term Frequency–Inverse Document Frequency) technique.

1.1 Introduction

In 1990, a telecommunications group in Japan introduced emoticons to motivate children to use their paging services. The word "emoticon" is derived from a Japanese expression meaning "picture character." It later spread to Western countries when users discovered a concealed keyboard feature on the Apple iPhone, originally designed for the Japanese market. Today, emoticons are commonly used across social media platforms to convey opinions and feelings. For example, companies can apply sentiment analysis to perform quantitative evaluations of their products without relying on traditional surveys that may inconvenience customers. Earlier approaches to sentiment analysis on Twitter mainly focused on classifying tweets into positive or negative categories, a method known as binary sentiment analysis. Emoticons, however, are visual symbols included within text and differ from processed textual data, as well as from emojis, which are facial expressions formed using ASCII characters. In modern times, emoticons are supported as an optional input language on many mobile devices and have become a key part of daily communication, helping to express emotions within text-based messages.

1.2 Background

Sentiment analysis, often referred to as opinion mining, is a method within natural language processing (NLP) used to classify text as positive, negative, or neutral. By evaluating the emotions expressed in written content, it is widely applied to analyze product reviews, survey responses, and customer feedback. Its use spans multiple areas, such as tracking social media activity, managing brand reputation, and improving customer satisfaction. For example, analyzing large volumes of product reviews can reveal important insights about features, quality, and pricing. Beyond simply determining polarity, sentiment analysis can also identify particular emotions (like anger, joy, or sadness), detect urgency (urgent versus non-urgent), and even infer user intent (interested or not interested).

1.2 Purpose

Sentiment analysis, also known as opinion mining, is widely used to evaluate movie reviews, product feedback, customer service responses, and reactions to various events. It helps determine whether a product or service is satisfactory, unsatisfactory, or preferred. In addition, it enables the assessment of public opinion about a particular event or individual by identifying the overall polarity of the text as positive, negative, or neutral. As a text classification technique, sentiment analysis organizes content into distinct emotional categories.

With the growth of social awareness, social media platforms such as Twitter have gained significant popularity. Twitter is a widely used networking platform where users share tweets about different events and topics. It acts as an open space for individuals to express their thoughts, opinions, and emotions freely. The increasing availability of affordable internet access and portable devices has made Twitter more accessible, enhancing its role in modern communication.

A large number of users rely on Twitter to discuss current events and share their perspectives. In today's digital age, people often use the platform to communicate their feelings and viewpoints, resulting in the generation of massive amounts of data. Even with the limitation of 140 characters per tweet, users effectively express their sentiments through concise wording. With hundreds of millions of active users, Twitter produces an enormous volume of tweets every day. These tweets frequently reflect users' opinions on specific subjects, making the platform a rich and easily accessible source of data for analysis.

1.3 Problem Statement

The primary problem addressed in this paper is divided into two key subtasks:

1. **Phrase-Level Sentiment Analysis:**

   This task focuses on evaluating the role of individual words or phrases within a message, particularly in social media content. The goal is to determine whether a specific term or expression conveys a positive, negative, or neutral meaning in its given context.

2. **Emotion-Level Sentiment Analysis:**

   This task aims to examine how text or emoticons influence the emotions or perceptions of others. Based on this evaluation, the sentiment is classified as positive, negative, or neutral, typically in relation to a particular event or situation.

1.4 Scope

An emoticon, for example :-) , is a simple representation of facial expressions in text form. It enables the writer to communicate feelings, moods, and emotions within a message while incorporating elements of nonverbal expression. This not only attracts the reader's attention but also improves their interpretation of the message. The development of emojis—originating from Japanese terms meaning "picture" and "character"—has further enhanced the ability to express ideas through modern digital communication. Emojis encompass a broad range of categories, such as celebrations, infrastructure and transportation, weather conditions, food and drinks, flora and fauna, along with emotions, actions, and activities.

This paper is structured as follows: Section 2 provides a comprehensive review of related research. Section 3 presents the emoji lexicon, and Section 4 explains the methodology used to analyze emotions expressed through emoticons. Section 5 outlines the experimental findings from the study, and Section 6 offers the final conclusions.

2. Related Research

In recent years, significant work has been carried out in the field of text-based sentiment analysis. Yogesh Chandra et al. [1] applied machine learning classifiers to perform sentiment analysis, where consumer tweets were categorized as positive or negative using polarity-based techniques along with deep learning models. Alzubaidi et al. [2] proposed a method for analyzing sentiment in social media posts by combining machine learning algorithms with emojis to classify emotions into neutral and negative categories. Lee et al. [3] introduced MultiEmo, a deep neural network that integrates emoji prediction with emotion detection tasks. Their model was evaluated using two datasets: a Twitter dataset for emoji prediction and the GoEmotion dataset for emotion classification. The findings indicated that MultiEmo surpassed existing approaches, demonstrating improved accuracy in predicting both emoji and emotional labels. Additionally, the study offered insights into the connection between emojis and emotions, enhancing the interpretability of the model's outputs and its decision-making process.

Subbashini L et al. [4] conducted a comprehensive review of the literature on opinion mining, discussing methods for representing and classifying knowledge derived from opinions, as well as techniques for extracting textual features from data containing uncertainty and noise. Mowlaei M.E. et al. [5] proposed a sentiment classification approach based on an aspect-oriented adaptive vocabulary. They introduced two strategies—genetic algorithms and statistical methods—to construct dynamic vocabularies for categorizing emotions. Kumar K.N. et al. [6] developed a dynamic dictionary capable of automatic updates, improving the classification of context-specific concepts. Zvarevashe K. et al. [7] utilized multiple vocabularies from different dictionaries to organize various components of their study.

Sentiment analysis has traditionally been applied across diverse sectors such as finance, aviation, hospitality, and healthcare. Mayur et al. [8] provided a detailed overview of sentiment analysis techniques, evaluating and comparing different methodologies to understand their strengths and limitations, while also identifying future research directions. C.H. Rayala Vinod Kumar et al. [9] focused on analyzing highly unstructured Twitter data, comparing their proposed approach with methods such as VLS (Vader Lexicon Sentiment) and convolutional neural networks (CNNs). Their research methodology explained the functioning of the algorithms, and results were presented based on both VLS and CNN techniques.

Although numerous studies have explored sentiment analysis, most have concentrated primarily on text mining. Emoticons and emojis, however, can complement textual content by reinforcing emotional meaning or even function independently. Comparative analysis indicates that incorporating both text and emoji-based features can enhance the overall effectiveness of sentiment classification methods.

3. Emoji Lexicon

Japanese mobile phone companies were among the first to introduce and popularize the use of emojis. Figure 1 illustrates a collection of these emojis.

Figure 1: Screenshot of Emoji's Template

A mobile service provider initially offered a set of 176 emojis for messaging purposes. Later, in 2010, Unicode 6.0 introduced a large number of emoji characters, which enabled their broader adoption across digital platforms. Subsequently, Unicode 9.0 expanded the set further, supporting a total of 1,126 distinct emoji characters.

Emojis are part of the broader Unicode standard, meaning they are stored and processed in the same way as other Unicode symbols. To increase user engagement, some websites design custom fonts that render emojis in a more visually attractive manner. However, depending on the programming language, encoding system, or documentation used, Unicode characters may appear in different formats or representations.

The way a sentiment analysis (SA) system collects and interprets data also depends on the tools and platforms used. As a result, researchers may encounter varying representations of emojis within datasets. For instance, emoji encoding in the Twitter API for R may differ from that used in Hadoop Flume. Therefore, using the correct Unicode representation is essential when searching for or analyzing specific emoji characters.

The emoji sentiment analysis model [10] introduces a technique for assigning sentiment scores to emojis based on the emotional ratings of associated text. Native speakers of the language were given plain text and asked to classify it as positive, negative, or neutral. The average sentiment score for each emoji was then calculated by aggregating the ratings from all corresponding texts. This process resulted in the creation of an emoji sentiment dictionary, which assigns values ranging from 0 to 1. This dictionary is useful for analyzing text and identifying emojis in data collected from microblogging and social media platforms.

id	user_description	user_location	text	user_name	user_created	user_followers	created	retweet_count	polarity	subjectivity
1	T H I C C latinas only	Trumbull, CT	â†™ #unpopularopinion #lol The inevitability of peace i	plive_calmer	2017-03-11T09:37:24	16	2019-12-07T14:13:54	0	0.225	0.35833333
2	p a r k s o s h	FA - RANDOM	RT @smolkjd: are we surprised he's trending on 3 differ	dyompolin	2018-12-16T12:38:58	1137	2019-12-07T14:13:54	0	-0.09107	0.75892857
3	â˜¯ï¸ People,Pets,Art	Tucson	RT @janewells: Jury foreman told us Unsworth did not p	sandi_childs	2013-11-16T11:53:08	4100	2019-12-07T14:13:54	0	0	1
4	Love to express my c	Islamabad, Pakista	@summayasaleem close call :p	iamwakas	2011-04-19T18:08:24	572	2019-12-07T14:13:54	0	0.75	1
5	[Roleplayer] What are you trying to fin		Bubu, @johnnysoeh meet my ooglie friends. They are a	peppermxnts	2019-11-30T15:00:58	13	2019-12-07T14:13:54	0	0.05	0.8875
6	http://GoldenRation	Around.	@madelyndewinter @NewYorker @hulu I donâ€™t hav	NautilusShell	2009-09-08T13:08:55	538	2019-12-07T14:13:54	0	-0.75	1
7	à¹€à¸›à¹‡à¸™à¸ªà¹ˆà¸		RT @GOT7Official: I GOT7 6th Generation Photo Teaser	_Support_YOU	2019-03-03T11:46:54	18	2019-12-07T14:13:54	0	0	0
8	ã€ ð—™ð—œð——–ð	US â€º JPN â€º KR â	@Lusinners message â€” Arcene.	rxndzvous	2019-07-16T15:03:44	108	2019-12-07T14:13:54	0	0	0
9	DAY6 ðŸŽ¸ BTSðŸ’£	iKONðŸ à¸£à¸±à¸	RT @caleridula: Jae : life is really hard sometimes, and i	lammyday6	2018-01-03T04:39:31	34	2019-12-07T14:13:54	0	0.252778	0.79722222
10	ì†Œë‹‰/ë°ë“œœìŠ¤íŽ	ë‚©ìž'ì´ˆ 124589ë²ˆ	RT @An_Rock_Sam: Third Krita Practice. I still like her.	EVA_Mk_1	2015-02-03T12:05:33	119	2019-12-07T14:13:54	0	0.25	0.5
11	Thalapathy Vijay loy	Chennai	ðŸŽ‰Active Thalapathy Fans Followers InvitedðŸ”¥	Itz_Starker	2017-04-22T06:24:39	889	2019-12-07T14:13:54	0	0	0
12		California, USA	RT @atensnut: To schedule a speaking - book signing	Stancy11382013	2018-08-18T17:45:05	1740	2019-12-07T14:13:55	0	0	0
13	Fashion . Radio Pers	Harvey, IL	RT @therealblair: Pt2 yo need a Grammy ðŸ˜,(he only 1	dsymoneallure	2015-01-13T07:27:58	101	2019-12-07T14:13:54	0	0	1
14	My sunshine Double B #BamBam , 2014.1		RT @BubblebamB: ðŸ½191207	bias_BamBamonly	2014-08-26T00:00:12	283	2019-12-07T14:13:55	0	0	0

Table 1: Emoji Lexicon Sample Entry

4. Methodology

The following sections present a detailed description of the step-by-step approach used for sentiment analysis. The dataset is collected using the Twitter API, then processed through a pre-processing stage, and finally analyzed using machine learning methods, as illustrated in Figure 2.

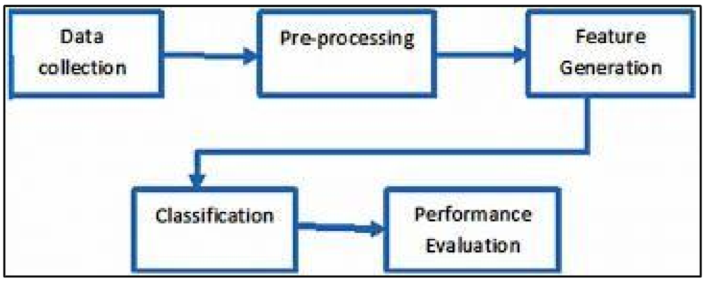

Figure 2: Flow diagram of Sentiment Analysis

4.1 Twitter API

Twitter offers two main types of APIs: REST and Streaming. The REST API includes both the standard REST interface and the Search API, which have developed separately over time. The main distinction between REST and Streaming APIs lies in how they deliver data. The Streaming API maintains a continuous connection and provides real-time updates, whereas REST APIs are rate-limited and designed for short-term use, allowing users to retrieve a limited amount of data within a specific period, such as per day. REST APIs are commonly used to access different kinds of Twitter data, including user profiles and notifications. However, Twitter makes historical data available only for a limited duration, typically about seven days, meaning REST access is restricted to tweets not older than one week. In contrast, the Streaming API delivers tweets as they are posted in real time, while the REST API is used to access previously collected or stored data.

4.2 Data Processing

Data undergo a series of steps as illustrated in Figure 3.

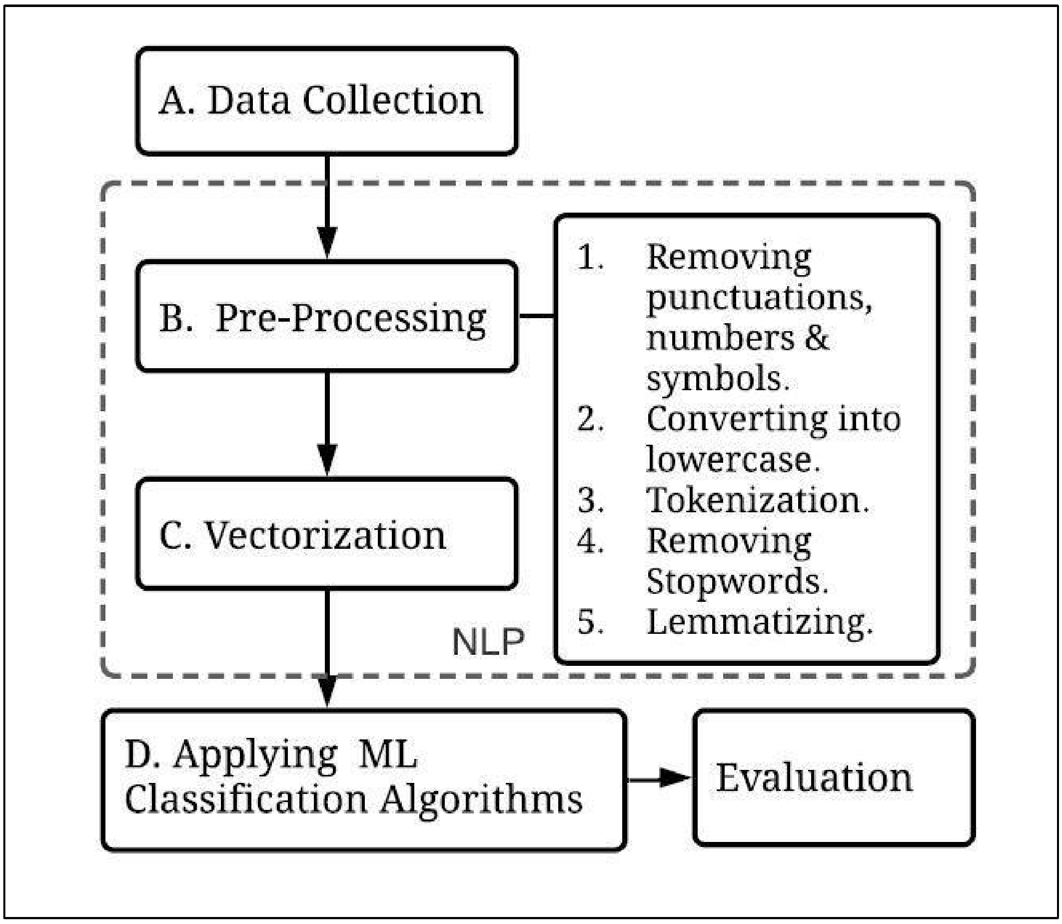

Figure 3: Step diagram

Data Aggregation:

In this stage, the information required for analysis is collected from multiple sources such as blogs and social networking platforms like Twitter and MySpace, depending on the application domain. In this study, data is gathered specifically through the Twitter API.

Preparation:

At this stage, the collected data is cleaned and structured for use in the classifier. Data cleaning includes removing unwanted symbols and extracting relevant phrases, along with standardizing text by converting uppercase letters to lowercase. Non-English content is excluded, and extra spaces and tabs are removed to improve data quality.

Training Data:

A well-structured dataset is created using widely accepted and transparent data selection methods to ensure reliability and consistency for training purposes.

Classification:

This is a key phase of the process. Based on the application requirements, either a Random Forest or Decision Tree algorithm is used for analysis. Once trained, the classifier is capable of identifying sentiments from continuous streams of tweets or other textual content.

Results:

The final outputs are displayed using visual tools such as graphs and charts. Before presenting the results, performance optimization and tuning are performed to ensure accuracy and clarity of the analysis.

4.3 Tokenization and Stemming

Tokenization is the process of breaking down large text into smaller units called tokens. Natural language processing (NLP) uses these techniques to develop applications such as text classification, intelligent chatbots, sentiment analysis, language translation, and others. Identifying patterns in text is essential for achieving these tasks, and tokens form the basic building blocks for such pattern recognition. They are also used as a foundation for techniques like stemming and lemmatization, which help in further text processing.

Stemming and lemmatization are methods in computational linguistics used to normalize words, phrases, and textual data for analysis. These techniques have been researched and applied in software systems since the 1960s. In this context, learners are introduced to stemming and lemmatization through a structured approach that includes their purpose, common evaluation methods, practical applications, and implementation steps using the Python NLTK library, a widely used toolkit for natural language processing.

4.4 Lemmatization

Lemmatization is the process of reducing words to their root or base form. The key difference between stemming and lemmatization is that lemmatization considers the context of the word and converts it into its meaningful base form, whereas stemming simply removes suffixes from words, which can sometimes produce incorrect or non-dictionary results.

For example, lemmatization correctly identifies the base form of words by using contextual understanding, while stemming may only strip endings without preserving meaning, which can lead to errors.

'Caring' → Lemmatization → 'Care'
'Caring' → Stemming → 'Car'

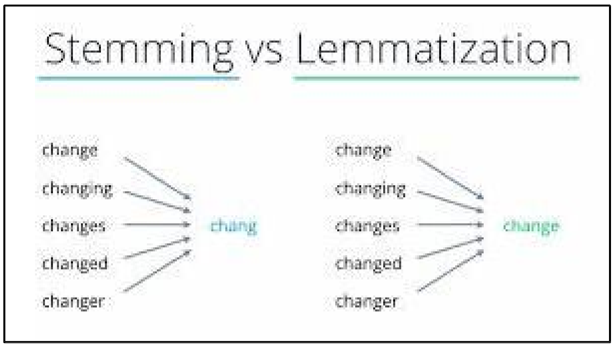

Figure 4: Example of stemming

4.5 Stop Words

Stop words are frequently occurring words like "if," "however," "we," "he," "she," and "they." These terms can often be removed from a text without significantly changing its meaning. Eliminating them may help improve model performance in some cases, although this is not always guaranteed.

4.6 MODEL

4.6.1 BOW Model

Bag of Words (BoW) is a technique, as shown in Figure 5, used to extract important features from text or emoticon-based documents. These extracted features are then used to train machine learning models. The BoW approach builds a vocabulary containing all unique words present in the reviews of a dataset. In simple terms, it represents a sentence as a collection of words without considering their order [11][12].

Bag of Words is widely used in natural language processing, information retrieval, and document classification tasks. At a general level, it involves the following steps.

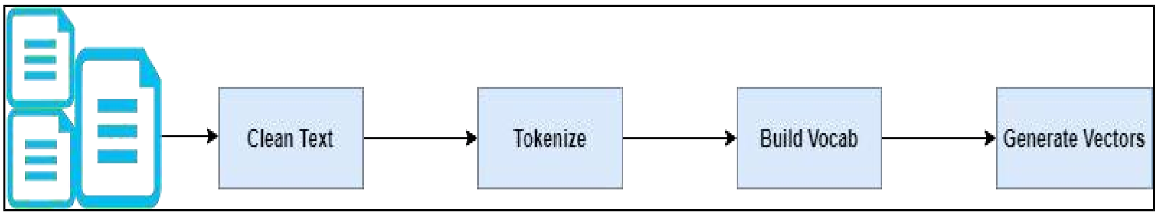

Figure 5: BOW Model

4.6.2 TF-IDF Model

TF-IDF stands for Term Frequency–Inverse Document Frequency, a statistical measure used to evaluate how important a word is within a corpus. The term frequency (TF) component indicates how often a word appears in a specific document within the dataset. It is calculated by dividing the number of times a word occurs in a document by the total number of words in that document. As the frequency of a word in a document increases, its TF value also increases. Each document in the corpus has its own TF value [11].

$f_{i,j} = {}_{i,j} / \sum {}_{i,j}$

The Inverse Document Frequency (IDF) component measures how rare or common a word is across the entire corpus. Words that appear in fewer documents are assigned higher IDF values, while commonly occurring words receive lower scores. It is defined as:

$() = \log (/ \quad {}_{t})$

The final TF-IDF score of a word in a document is obtained by multiplying TF and IDF values:

${}_{I,j} = f_{i,j} \times \log (/df_{i})$

Where tfi,j= count of occurrences of i in j

dfi= number of documents containing the ith word

N= total number of documents

4.7 Classifier

We are utilizing two classifiers for the empirical study. The subsequent subsections provide concise explanations of the algorithms.

4.7.1 Decision Tree classification

Decision trees are a type of supervised learning algorithm used for classification tasks. They belong to the family of information-based learning methods and rely on different strategies for acquiring information during the training process. Decision trees can handle problems where both input features and output targets are either continuous or categorical.

The primary goal of a decision tree is to identify the most informative features that provide maximum information about the target variable. The dataset is repeatedly split based on these feature values, resulting in subsets that become increasingly homogeneous with respect to the target class. Features that produce the purest splits are considered the most informative. This recursive splitting continues until a stopping condition is reached, forming terminal nodes known as leaf nodes. These leaf nodes represent the final predictions for new input instances provided to the model.

Since the model learns the underlying structure of the training data, it can predict the target class of unseen data under certain assumptions. A decision tree is composed of a root node, internal nodes, and leaf nodes, which are connected through branches. Decision trees can also be categorized based on whether the output variable is continuous (such as predicting real estate prices) or categorical (such as classifying animal species).

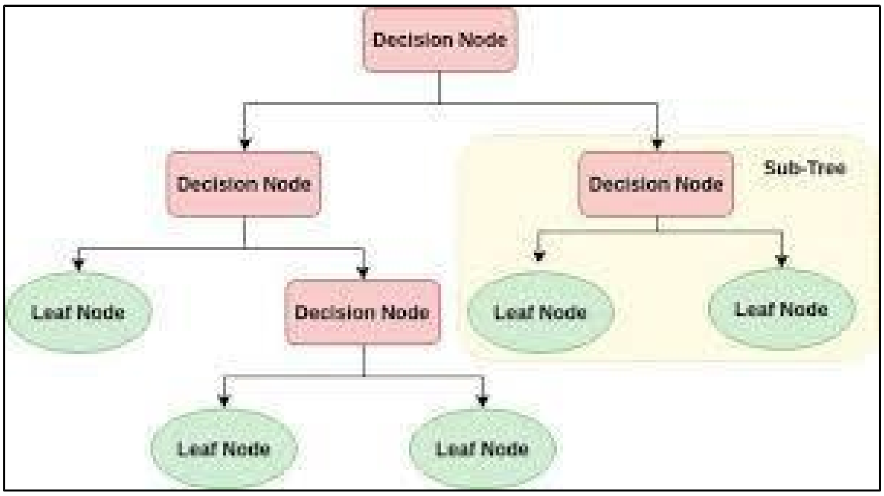

Figure 6: Decision Tree classification

4.7.2 Random Forest Classification

Random Forest is a widely used supervised learning algorithm in machine learning that can be applied to both classification and regression tasks. It is based on the concept of ensemble learning, where multiple models are combined to solve complex problems more effectively and improve overall performance.

As the name implies, a Random Forest is made up of several decision trees trained on different subsets of the dataset. Instead of relying on a single decision tree, the algorithm aggregates the outputs of all trees. For classification tasks, it uses a majority voting mechanism, while for regression tasks it typically averages the predictions of all trees to produce the final result.

Increasing the number of trees generally improves the model's accuracy and helps reduce the risk of overfitting. As a result, Random Forests are capable of handling diverse datasets efficiently while producing stable and reliable predictions.

The figure below illustrates the working process of the Random Forest algorithm.

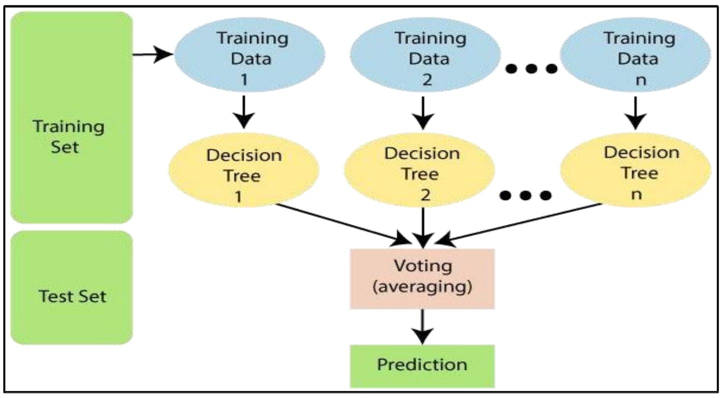

Figure 7: Tree structure of RF

5. Results and Analysis

A pie chart can be used to clearly represent the final results by showing how hashtags are distributed across different sentiment categories, such as positive, negative, and neutral. Hashtags that do not have any assigned sentiment value are treated as null.

For prediction purposes, both Decision Tree and Random Forest algorithms are applied. The findings of the study show that the Random Forest model provides slightly more accurate results compared to the Decision Tree model. The accuracy of each model is then evaluated, where the Decision Tree achieves an accuracy of 0.71, while the Random Forest achieves an accuracy of 0.72.

In addition, the emotional content of text is analyzed by comparing it with the emojis used within the same text. In some cases, the emotions expressed in the text match the associated emojis, while in other instances they may differ.

5.1 Results for BOW Model

The output of the Bag of Words (BoW) model is presented in Table 2. In this representation, a vector is generated to show whether each word in a sentence is a commonly occurring term or not. If a word is identified as a common term, it is given a value of 1; otherwise, it is assigned a value of 0.

	32	33	34	35	36	37	38	39	40
251	0	0	0	0	0	0	0	0	0
252	0	0	0	0	0	0	0	0	0
253	0	0	0	0	0	0	0	0	0
254	0	0	0	0	0	0	0	0	0
255	1	0	0	0	0	0	1	0	0
256	0	0	0	1	0	0	0	0	0
257	0	0	0	0	0	0	0	0	0
258	0	0	0	0	0	0	0	0	0
259	0	0	0	0	0	0	0	0	0
260	0	0	0	0	0	1	0	0	0
261	0	0	0	0	0	0	0	0	0
262	0	0	0	0	0	0	0	0	0
263	0	0	0	0	0	0	0	0	0
264	0	0	0	0	0	0	0	0	0
265	0	0	0	0	0	0	0	0	0
266	0	0	0	0	0	0	0	0	1

Table 2: Result of BOW Model

5.2 Result of TF-IDF model

In Table 3, a term frequency matrix is constructed in which each row represents a document and each column represents the unique terms present across the documents. All occurrences of each word are identified within every text.

After this, the inverse document frequency (IDF) is calculated using the standard formula. Once both TF and IDF values are obtained, the TF matrix is multiplied by the IDF values to produce the final TF-IDF representation.

The resulting processed text is then ready to be used as input for a machine learning algorithm.

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0	0	0	0	0	0	0	0	0	0	0.0869361	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0.0314157
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0.0781397	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0.0342694	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0.0917886	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0.13075	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0.0341868	0	0.030967	0	0	0	0	0
14	0	0	0.0949551	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0.103016	0	0	0

Table 3: Result of TF-IDF model

5.3 Confusion Matrices for DT classifier

A confusion matrix is generated for the data in the DT classifier, as illustrated in Table 4.

	0	1
0	74	23
1	35	68

Table 4: Confusion matrix for Decision Tree

The initial column in the dataset contains the value of (74+35) = 109. The total value of the second column in the dataset is (23+68) = 91. When predicting the first column as the second column, more errors were observed compared to the reverse scenario.

5.4 Confusion Matrices for Random Forest classifier

Table 5 displays the confusion matrix that we generate for our data in the Random Forest classifier.

	0	1
0	87	10
1	46	57

Table 5: Confusion matrix for Random Forest

The initial column of the dataset displays the sum value of (87+46) which equals 133. The total value of the second column in the dataset is (10+57) which amounts to 67.

5.5 Accuracy of DT and RF Algorithm

Accuracy is calculated using the formula (TP + TN) / (TP + TN + FP + FN).

Table 6 presents a comparison of the results obtained from the two algorithms. The Decision Tree model achieves an accuracy of 71.00, while the Random Forest model achieves an accuracy of 72. In terms of performance metrics, the Decision Tree shows precision and recall values of 74.725 and 60.019, respectively. On the other hand, the Random Forest model records precision and recall values of 85.074 and 55.339, respectively.

	Decision Tree	Random Forest
Accuracy	71	72
Precision	74.725	85.074
Recall	66.019	55.339
F1 Measure	70.102	67.058

Table 6: Comparative chart of algorithm accuracy

5.6 Outcomes of analyzing the sentiment of tweets

5.6.1 Weakly Positive events with Emoji

The primary objective of this study was to promote a positive atmosphere during the "Tokyoolympics" by using emoticon characters. For data collection, tweets containing the hashtag "#Tokyoolympics" were extracted, resulting in approximately 1000 samples. The collected data was then analyzed and visualized using a pie chart to provide a clear overview of the results.

As shown in Figure 7, different colors were used to represent the distribution of sentiment categories, including positive, weakly positive, strong positive, negative, weakly negative, strong negative, and null sentiment hashtags. Overall, the analysis indicated that the majority of users expressed a weakly positive sentiment.

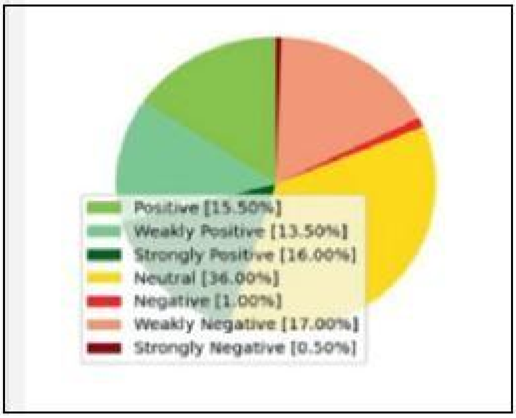

Figure 7: Pie Chart of the analyzed tweets

From the pie chart, it can be observed that 15.50% of users expressed a positive sentiment, while 13.50% showed a mildly positive opinion. In addition, 16.00% of individuals reported a strongly positive sentiment.

On the other hand, 1.00% of users expressed a negative sentiment, whereas 17.00% indicated a slightly negative view. Furthermore, 0.50% of respondents showed a strongly negative sentiment. Finally, 0.50% of the users were classified as neutral.

5.6.2 Neutral events with Emoji

We focused on the use of emoticon characters during the “Cheers4India” events, particularly those expressing positive emotions, since discussions related to “Cheers4India” often reflect negative sentiments among users. For data collection, tweets containing the hashtag “#Cheers4India” were extracted, resulting in approximately 1000 samples.

The analysis of the data is presented in a pie chart, which provides an overall summary. As shown in Figure 8, the general report indicates a neutral sentiment.

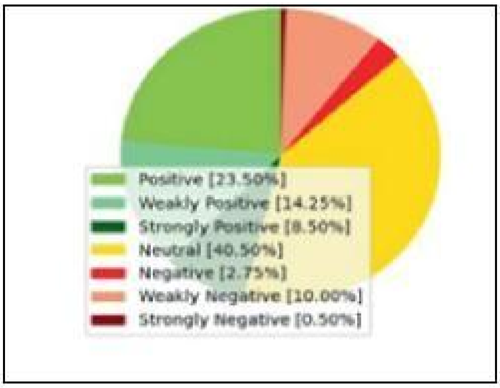

Figure 8: Pie Chart of the analyzed tweets

Sentiment hashtags are represented visually in the pie chart, where each category is assigned a distinct color, as shown in Figure 8. According to the chart, 10.00% of users expressed a weakly negative sentiment, while 2.75% showed a negative sentiment. In comparison, 23.50% of individuals exhibited a positive sentiment, and 14.25% demonstrated a weakly positive sentiment. Notably, a significant portion, 40.50%, was classified as having a neutral sentiment.

Two pie charts were analyzed based on the topics “Tokyoolympics” and “Cheers4India.” Figure 7 shows that weakly positive sentiment was dominant, whereas Figure 8 indicates that neutral sentiment was more prevalent.

5.7 Result and Analysis

Figure 9 shows a bar chart illustrating the relationship between the emotions expressed in textual content and the emojis used alongside it. It also highlights cases where the sentiment conveyed by the text does not match the emotion represented by the emojis. The observed ratio of matching emotions to mismatched emotions is 8.2:5.0.

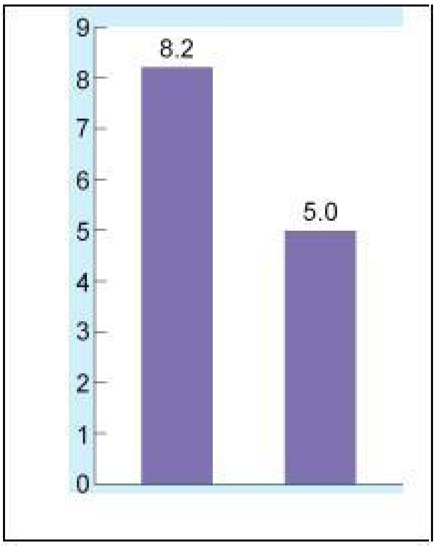

Figure 9: Result for compare the emotions of a text & its emojis which is present in that text

6. Conclusion

This study investigates the use of emoticon characters on social media platforms, especially within informal online communities. It empirically explores how emoticons influence text mining outcomes and how they relate to expressed emotions. The analysis covers a range of globally trending topics to identify whether differences exist between emotions conveyed through emoticons and those expressed in accompanying text.

The results show that including emoticons in emotion analysis improves the overall sentiment scoring. Although emoji characters can represent both positive and negative emotions, the findings suggest that emoticons tend to enhance the expression of positive sentiment and lead to higher emotional scores compared to negative sentiment. For data collection, a crawler was used to extract tweets from Twitter, which were then processed through feature removal during the retrieval phase.

Sentiment analysis has been applied for more than a decade, and it is now widely used by organizations as a support tool for decision-making and strategic planning. Its growing importance is also supported by advancements in data storage, accessibility, and processing enabled by big data technologies. Emoji analysis provides a new dimension to traditional text-based sentiment analysis. Since emojis are often used alongside text to add emphasis or context, examining the relationship between textual content and emojis is essential for a more complete understanding of emotions in online social networks.

Disclosure statement: The authors report there are no competing interests to declare.

References

[1] Yogesh Chandra, Antoreep Jana, “Sentiment Analysis using Machine Learning and Deep Learning”, 7th International Conference on Computing for Sustainable Global Development (INDIACom),2020.

[2] M. SalehAlzubaidi and F. Bourennani, “Sentiment analysis of tweets usingemojis and texts,” in ACM International Conference Proceeding Series, ACM, Mar. 2021, pp. 111–114, doi: 10.1145/3459955.3460608.

[3] S. E. Lee, D. Jeong, and E. Park, “MultiEmo: Multi-task framework for emoji prediction,” Knowledge-Based Systems, vol. 242, Art. no. 108437, Apr. 2022, doi: 10.1016/j.knosys.2022.108437

[4] Subhashini L, Li Y, Zhang J, Atukorale AS and Wu Y, “Mining and classifying customer reviews: a survey”, Artificial intelligence Review 54:6343–6389, 2021.

[5] Mowlaei ME, Abadeh MS, Keshavarz H, “Aspect-based sentiment analysis using adaptive aspect based lexicons”, Expert System Application, 148:113234, 2020.

[6] Kumar KN, Uma V, “intelligent sentiment-based lexicon for context-aware sentiment analysis: optimized neural network for sentiment classification on social media”, Journal of, 2021.

[7] Zvarevashe K, Olugbara OO (2018),”A framework for sentiment analysis with opinion mining of hotel reviews”, In: 2018 Conference on information communications technology and society (ICTAS).IEEE, pp 1–4

[8] MayurWankhade, Annavarapu Chandra SekharaRao, ChaitanyaKulkarni,” A survey on sentiment analysis methods, applications, and challenges”, Artificial Intelligence Review ,2022.

[9] CH. RayalaVinod Kumar, D. LalithaBhaskari, P. SrinivasaRao, “VLS Algorithm: A New Approach to Sentiment Analysis”, International Journal of Technology and research, Vol 9, Issue 05, MAY 2020, ISSN 2277-8616.

[10] Kralj Novak, P., Smailović, J., Sluban, B., and Mozetič and I.,” Sentiment of emojis”, PLOS One, 10(12), 2015.

[11] M. Fernández-Gavilanes, E. Costa-Montenegro, S. García-Méndez, F. J. González-Castaño, and J. Juncal-Martínez, “Evaluation of online emoji description resources for sentiment analysis purposes,” Expert Systems with Applications, vol. 184, Art.no. 115279, Dec. 2021, doi: 10.1016/j.eswa.2021.115279.

[12] V. A. Pfeifer, E. L. Armstrong, and V. T. Lai, “Do all facial emojis communicate emotion? The impact of facial emojis on perceived sender emotion and text processing,” Computers in Human Behavior, vol. 126, p. 107016, Jan. 2022, doi: 10.1016/j.chb.2021.107016.