Social Network Analysis and Mining (2022) 12:137
https://doi.org/10.1007/s13278-022-00961-1

ORIGINAL ARTICLE

A self-attention hybrid emoji prediction model for code-mixed language: (Hinglish)

Gadde Satya Sai Naga Himabindu[1] · Rajat Rao[1] · Divyashikha Sethia[1]

Received: 19 June 2022 / Revised: 27 July 2022 / Accepted: 21 August 2022 / Published online: 14 September 2022

Abstract

Emojis are an essential tool for communication, and various resource-rich languages such as English use emoji prediction systems. However, there is limited research on emoji prediction for resource-poor and code-mixed languages such as Hinglish (Hindi + English), the fourth most used code-mixed language globally. This paper proposes a novel *Hinglish Emoji Prediction (HEP)* dataset created using Twitter as a corpus and a hybrid emoji prediction model *BiLSTM attention random forest (BARF)* for code-mixed Hinglish language. The proposed BARF model combines deep learning features with machine learning classification. It begins with BiLSTM to capture the context and then proceeds to self-attention to extract significant texts. Finally, it uses random forest to categorize the features to predict an emoji. The self-attention mechanism aids learning since Hinglish, a code-mixed language, lacks proper grammatical rules. The combination of deep learning and machine learning algorithms and attention is novel to emoji prediction in the code-mixed language(Hinglish). Results on the HEP dataset indicate that the BARF model outperformed previous multilingual and baseline emoji prediction models. It achieved an accuracy of 61.14%, precision of 0.66, recall of 0.59, and F1 score of 0.59.

Keywords Emoji prediction · Hinglish · Code mixed · Deep learning · Hybrid model

1 Introduction

As the Internet and social media platforms have grown in popularity, individuals have become used to expressing various emotional tendencies and feelings through social media platforms. As a new visual language, emojis are essential for conveying emotion and amplifying the visual impact of short text messages (Barbieri et al. 2016). To communicate on social media platforms, people usually use informal short text messages to give them a personal touch, which results in ambiguity. Emojis reduce this ambiguity in short text messages by conveying the intended tone. For example, *see you tomorrow* , here tone is not clear, so without context, it is on the reader how he takes it. However, the use of emojis can reduce ambiguity in it. *see you tomorrow*😁 conveys excitement and joy while *see you tomorrow*😡 conveys anger. This ability of emojis to convey tone and emotion in short text messages made them extremely popular.

Emojis' extensive use and popularity have created a new language of symbols that is constantly expanding, with new emojis arising with varying meanings day by day. Although hundreds of emojis are available, users cannot efficiently use them due to the time-consuming task of selecting an emoji from hundreds of options. This problem has prompted research into emojis and their relationship to text (Aoki and Uchida 2011). With the advancement in natural language processing (NLP) and its applications, emoji prediction has become one of the most exciting social media research topics. The emoji prediction task attempts to predict relevant emoji(s) that can fit the context based on the text input.

However, the increasing popularity of social media platforms in linguistically distinct demographic regions proposes a new challenge for emoji prediction due to the code-mixing of languages. Code mixing is the mixing of two or more languages while communicating. Native speakers and code-mixers tend to change the script of their native language. Users write original code-mix text in Roman script,

✉ Gadde Satya Sai Naga Himabindu
gssnhimabindu@gmail.com

Rajat Rao
rajatrao006@gmail.com

Divyashikha Sethia
divyashikha@dtu.ac.in

[1] Department of Computer Engineering, Delhi Technological University, Delhi 110042, India

resulting in no fixed grammatical norms and phonetic differences due to regional influence. It becomes quite challenging in the Indian context as there are 22 official languages and numerous phonetic variations due to diverse cultural and regional influences. Hinglish is an important code-mix language (Parshad et al. 2016). Previous research focuses on semantic analysis, but there is a need to have an emoji prediction model for Hinglish. This paper attempts to explore emoji prediction for the Hinglish language for the first time to the best of our knowledge.

The main contributions of this paper are summarized as follows:

- *Hinglish Emoji Prediction (HEP) dataset* This work creates a new HEP dataset,[1] which is a collection of Hinglish tweets having emojis from Twitter. The annotated Hinglish (Hindi–English) code-mixed dataset for emoji prediction can enable future researchers to contribute to this domain.
- *BiLSTM attention random forest (BARF) model* Proposal of a hybrid BARF emoji prediction model for Hinglish, one of the most used code-mixed languages globally. Hinglish presents challenges with no fixed grammatical rules and spelling variations due to Roman script. “Jhooth bolna galat baat hai,” for example, implies that lying is wrong. Due to many possible pronunciations, the term “jhooth” can have several phonetic variants such as “jhuth,” “jhoot,” or “jhut.” These different terms may have distinct meanings in different contexts, resulting in ambiguity. The proposed model uses BiLSTM, self-attention, and random forest to predict our dataset’s top 40 frequently used emojis. Deep learning methods, along with machine learning methods, enhance prediction accuracy. Deep learning methods are good at extracting features and learning semantic expressions, which were otherwise done manually in traditional methods. BARF uses self-attention to aid learning as it helps to overcome the problem of no fixed grammatical rules in Hinglish.
- *Comparison of BARF with other models* Compared our model with various machine learning models, including Naive Bayes and random forest, as well as deep learning models such as CNNs, LSTMs, and bidirectional LSTMs (with and without attention).

The paper is structured as follows: Section 2 contains background information and relevant work on this topic. Section 3 describes the approach used in this paper to conduct the tests, including pre-processing, embeddings, and models. Section 4 lists the experimental parameters used to reproduce the work. Section 5 details the outcomes and contains conclusions based on the data.

2 Related work

Emojis are a type of ideographic character widely used on social media platforms. They help express emotions intuitively and alter the overall semantics of short text messages. Emojis not only reduce ambiguity in short text messages but also help to convey the tone. Therefore, emojis play an important role in short text messages. However, emojis do not have grammar rules, so their usage is subjective. Emojis have been the subject of a number of academic research, as discussed in the following section. Initially, the focus was on the descriptive analysis of the usage of emojis, and they were explored as emotional annotations in plain text (Vidal et al. 2016). With the advancements in natural language processing (NLP) and further research on emojis, emoji prediction received more attention.

2.1 Embeddings and emoji prediction

Barbieri et al. (2016a) used distributional embeddings to study emoji semantics. Using the Skip-gram model, the authors trained emoji embeddings on 100 million English Twitter tweets. The model indicated how embeddings could improve pair similarity and clustering accuracy. The proposed methodology was further used in different languages to study emoji usage (Barbieri et al. 2016, 2016b). Pohl et al. (2017) proposed a similar embedding model for computing emoji similarity. The model allowed to place related emojis together, thus making it easy for the user to use them. The neural embedding model gave a good performance for computing similarity between emojis. Eisner et al. (2016) proposed emoji2vec, pre-trained embeddings for emojis. Embedding methods like skip-gram and word2vec are not efficient in the case of infrequent emojis. The model directly trained embeddings from the Unicode description of emojis. Emoji2vec outperforms the neural embedding model proposed by Barbieri et al. (2016a). Wijeratne et al. (2017) further improved the embedding model for emojis by incorporating different representation methods, namely emoji description, emoji definition, and emoji sense labels. The authors created the EmoSim508 dataset, which is publicly available, consisting of 508 emoji pairs for performing emoji similarity calculation tasks.

2.2 Deep neural network and emoji prediction

For modeling emoji semantics neural models are used because of their efficiency in learning features. The majority of the neural models use word embeddings generated from FastText, word2vec, or GloVe. Xie et al. (2016) explored emoji recommendation tasks in cross-conversational systems. They proposed a hierarchical long short-term memory

[1] https://github.com/Himabindugssn/HEP.

(LSTM) network to model contextual information for the emoji prediction task. Their proposed model outperformed other LSTM models, namely Single-LSTM and Flattened-LSTM. Barbieri et al. (2017) first proposed an automatic method to predict emojis for given short text input. They proposed a bidirectional long short-term memory (BiLSTM) model (Graves and Schmidhuber 2005) with word and character-based representations (Ling et al. 2015) which outperformed bag-of-words (BOW) baseline and skip-gram vector average-based baseline model.

2.3 Attention mechanism and emoji prediction

Felbo et al. (2017) used a variant of the LSTM neural network along with an attention mechanism (Yang et al. 2016) to detect sentiment, emotion, and sarcasm through emojis. Based on this model, Barbieri et al. (2018) proposed a label-wise attention mechanism. It improved the robustness of Recurrent Neural Network(RNN) models in datasets with unbalanced distributions, as ltekin and Rama (2018) showed that they, do not perform better than SVMs on the emoji prediction task.

Guibon et al. (2018) proposed an emoji prediction model which could be trained and tested on actual data from text messaging applications. The proposed multi-label random forest (RF) classifier model with BOW/character representation outperformed BiLSTM networks.

For the multi-label emoji prediction task, Wu et al. (2018) proposed a Hierarchical neural model with an attention mechanism. The proposed model used Convolutional Neural Networks (CNN) to learn hidden word representations, a CNN and LSTM-based word encoder to learn sentence representations, and an attention mechanism. The proposed model outperformed the Support Vector Machine (SVM), CNN, and Hierarchical LSTM model in the emoji prediction task.

2.4 Multilingual emoji prediction

Barbieri et al. (2018) introduced a multilingual emoji prediction task at SemEval in which there were two subtasks, one for emoji prediction in English and one for Spanish. Out of the 49 participating teams, 22 teams submitted for Spanish subtask. The majority of the top-performing teams preferred CNN or LSTM-based neural networks. Hence they are considered baseline models for testing the validity of the proposed model. However, Tubingen-Oslo ltekin and Rama (2018) performed best in both tasks, which used an SVM classifier with bag-of-n-grams features. Performance metrics of these models can be seen in Tables 1 and 2 where BiLSTM + Attention represents the system proposed by NTUA-SLP Baziotis et al. (2018) and SVM with n-gram is taken from the system proposed by ltekin and Rama (2018).

On the similar lines to SemEval, EVALITA 2018 evaluation campaign proposed an emoji prediction task for Italian Language Ronzano et al. (2018). Five different teams made 12 submissions in total for this task. Most top-performing teams employed neural architectures and achieved good accuracy, especially using the BiLSTM model.

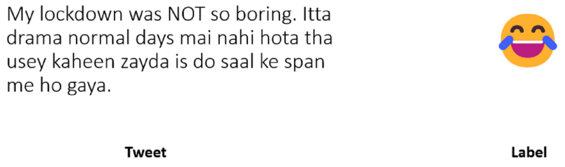
For the first time, Tomihira et al. (2018) explored the emoji prediction task for the Japanese language. They proposed an encoder–decoder with attention that outperformed CNN and RNN-based models. Another study by Tomihira et al. (2020) explored the Bidirectional Encoder Representations from Transformers (BERT) model, which performed better than the conventional model like FastText, CNN, the Attention BiLSTM. They also compared the Japanese BERT model with the English BERT model, though it scored less than the English BERT score, which may be lower due to the Japanese dataset.

Choudhary et al. (2018) addressed the issue of primarily ignored resource-poor languages for emoji prediction and sentiment analysis by creating a corpus for Hindi, Bengali, and Telugu. Choudhary et al. (2018) introduced Classification of Emojis using Siamese Network Architecture(CESNA), a twin BiLSTM network-based emoji prediction model for resource-poor languages. They trained Hindi and Telugu (resource-poor language) and English and Spanish(resource-rich language) simultaneously in the same emoji space.

Liebeskind and Liebeskind (2019) explored the emoji prediction task for the Hebrew language, considering it a single-label classification problem. They investigated different dimension reduction methods used to associate similar words to similar vector representations against machine learning algorithms. They found that common word embedding dimension reduction methods are not optimal. They showed that n-grams and character n-grams representations significantly outperform other vector representations for emoji prediction tasks in Hebrew.

Peng and Zhao (2021) explored emoji prediction for the Chinese language. They proposed an encoder–decoder model that utilizes attention for emoji prediction. BiLSTM-CNN was used for the encoder to understand the input sentence's global and local semantic information. Simultaneously the attention mechanism increases the weight of words with a significant contribution, which helps to improve prediction accuracy. The decoder used two RNNs in different directions to decode and predict the emojis.

While there are many studies on resource-rich languages, the field of code-mixed (Hinglish) text is still relatively new and unexplored. Most past research is on monolingual datasets since a large corpus of annotated data is readily available. The fundamental challenge in dealing with code-mixed situations is a lack of well-labeled datasets, besides ambiguity in code-mixed language. With the growing number



My lockdown was NOT so boring. Itta drama normal days mai nahi hota tha usey kaheen zayda is do saal ke span me ho gaya.

Fig. 1 Example of text and corresponding emoji from our HEP dataset

of non-native English speakers on social media, sentiment analysis, hate speech detection on regional languages, and code-mixed data have gained traction. According to a thorough review of data from English–Hindi bilingual Facebook users (Bali et al. 2014), 17.2% of all postings, accounting for about one–fourth of the words in their dataset, exhibited some code-mixing.

While sentiment analysis (Vijay et al. 2018), Hate speech detection (Mathur et al. 2018) for Hinglish has been explored in the past, but emoji prediction has not been explored yet. Hence, this paper investigates the emoji prediction for Hinglish code-mixed language.

3 Proposed methodology

3.1 Dataset

This work proposes creating a new Hinglish Emoji Prediction (HEP) dataset for the Hinglish emoji prediction task. It comprises of retrieval of tweets containing Emojis using the Twitter API, with Twitter serving as the corpus.

The HEP dataset comprises plain text as input and emojis as labels, as shown in Fig. 1. It uses the emojis extracted from tweets as labels since manual annotation can introduce bias. The former method helps us to understand the real-world usage of emojis in a better way.

Out of 2,62,408 tweets, there are 86,072 usable tweets after filtering. The retrieved tweets are from a radius of 1000 kilometers, using Delhi(India) as the center, as the probability of obtaining Hinglish tweets is higher in this region.

The annotated tweets are divided into three categories: English, Hinglish, and Others (foreign script tweets, tweets of languages other than English and Hinglish). The following is the distribution of the three classes: Tweets in English-14,006, Hinglish-51,756, and Others-20,310.

The HEP dataset comprises tweets extracted using various hashtags such as Demonetization, NamasteTrump, Election 2019, Olympic Games, and Farmer Protest related to critical events that capture various emotions, thereby helping us maintain a balanced dataset.

Hindi keywords like aapka, accha, ajeeb, batao, chutti, dard, gussa, haal, hai, hasna, humari, koon, pareshan, pyaar, shaadi, suno, theri, wale, yaad, yaar,hai yaar, hamari, accha ,ajeeb, padhai, chup, halat, etc. are also used for extracting tweets.

3.2 Preprocessing

The data cleaning process removes all the unnecessary details as they are not aiding in learning, with the following steps:

- Removal of all emojis from the text.
- Removal of URLs from the text.
- Removal of all user tags, hashtags from the text.
- Removal of extra spaces from the entire text.
- Change all text characters to lower case.

It retains stop words since users switch from English to specific Hindi terms to enhance the tweet's content and personalize it. The removal of these words results in the loss of vital information. The model treats Hinglish as a stand-alone language under the proposed method without considering grammatical rules. The generated tweets are treated as individual words and phrases to generate word vector representation.

3.3 Proposed model

For emoji prediction, this paper proposes a BiLSTM attention random forest (BARF) model that employs BiLSTM, self-attention, and random forest, as shown in Fig. 2.

3.3.1 Embedding layer

For input, the proposed model uses FastText pre-trained word embeddings. FastText is an open-source framework that enables the learning of unsupervised text representations and supervised text classification. The benefit of FastText over Word2Vec is that the former treats a single word as a character-level n-gram, whereas the latter treats a single word as a single vector in space.

The model must learn each n-gram with a unique representation so that uncommon words can share n-grams with other words. For out-of-vocabulary terms, which are present in the testing set but not in the training set, the model learns the representation of each n-gram in an earlier stage and represents a missing word as a concatenation of vectors of n-grams. Model experiments with alternative word embedding approaches such as Word2Vec, Glove, and FastText. The results indicate that FastText performs better since it can overcome vocabulary errors

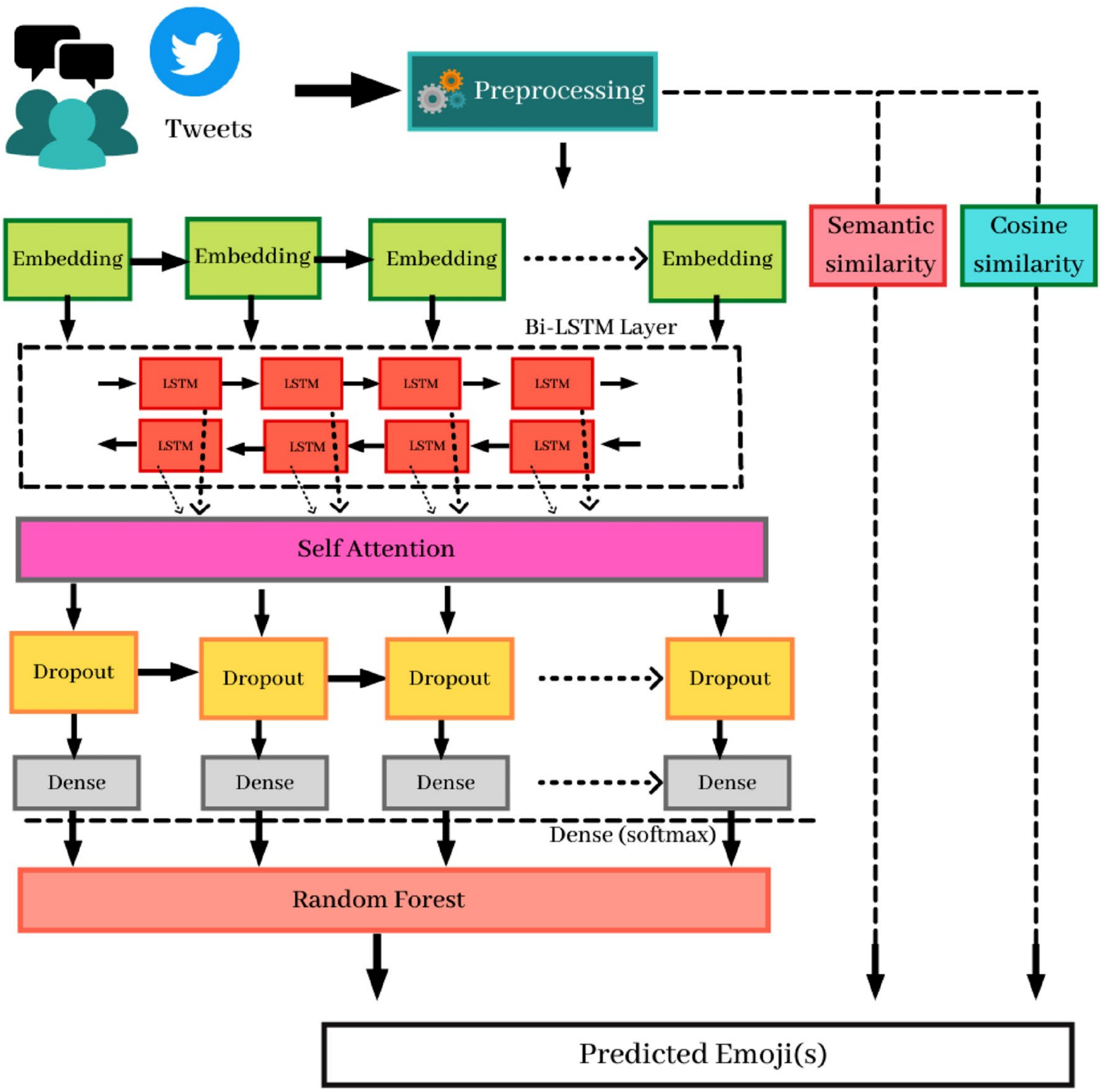


Fig. 2 Proposed Hinglish emoji prediction model

and construct representations for unusual words common in code-mixed languages.

3.3.2 BiLSTM layer

The arrangement of gates in LSTM benefits from sustaining long-term dependencies. BiLSTM has an additional advantage over LSTM in that it can traverse the input data from right to left and vice versa, extracting both past and future contexts. Hence, the model adopts BiLSTM as the first layer to capture all long-term dependencies while considering past and future contexts. This information is subsequently passed on to the attention layer as shown in Fig. 3.

3.3.3 Attention layer

The self-attention mechanism is used to capture significant parts of the text. The proposed model uses attention mechanism proposed by Bahdanau et al. (2014). The self-attention mechanism learns the correlation between the current word and the other parts of the sentence. The self-attention mechanism increases the weight of words with significant contributions, which improves prediction accuracy. This layer hence captures the vital part of the sentence. Random forest then classifies these features. It can efficiently deal with the high dimensional data and, at the same time, avoids overfitting as it bags various decision trees. The outputs are then processed by a voting mechanism, finally resulting in the prediction of an emoji.

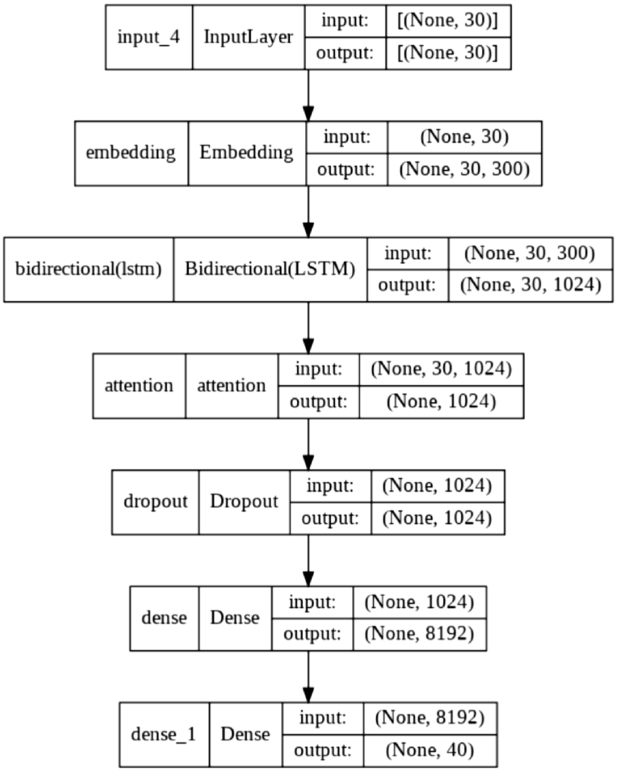


Fig. 3 BiLSTM + Attention (Partial BARF Model Architecture)

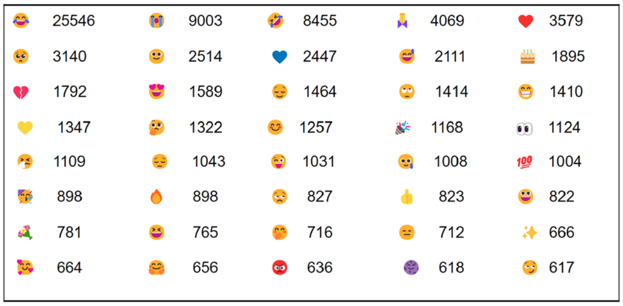
3.3.4 Random forest classifier

The information from the first fully connected layer of the BiLSTM attention model is extracted and sent to the random forest classifier to classify the features into one of the 40 emoji labels.

Random forest is a bagging method that employs the Ensemble Learning approach. It constructs as many trees as it can on the subset of data and then merges the results of all the trees. As a result, it decreases the overfitting problem in decision trees and the variance, thereby improving the accuracy. It is rapid in both model training and assessment, is resistant to outliers, captures complicated nonlinear relationships, deals with class imbalance data, and delivers competitive results for high dimensional data (Hastie et al. 2009), (Han et al. 2021). It has also been demonstrated to deal with problems caused by limited sample sizes (Qi 2012).

3.3.5 Cosine and semantic similarity

Cosine Similarity and Semantic algorithms are employed to predict emojis, which can be easily identified from their meaning like 🍰cake, 🍃leaf, and 🔥fire. Algorithms use an emoji dictionary consisting of 1700 emojis and their description through which emojis can be predicted based on similarity.



😂	25546	😭	9003	🤣	8455	🙏	4069	❤️	3579
🥺	3140	🥹	2514	💙	2447	😅	2111	🎂	1895
💔	1792	😍	1589	😌	1464	🙄	1414	😁	1410
💛	1347	🤔	1322	😊	1257	🎉	1168	👀	1124
🤧	1109	😔	1043	😜	1031	😥	1008	💯	1004
🥳	898	🔥	898	😒	827	👍	823	😃	822
💐	781	😆	765	🤭	716	😑	712	✨	666
🥰	664	🤗	656	😡	636	🌚	618	😏	617

Fig. 4 Selected emojis and their distribution

4 Experiments and results

The proposed BARF model predicts emoji out of the selected 40 most used emojis in the HEP dataset as shown in Fig. 4. For the proposed model, the input word length is set to 30 because the average length of a tweet is around 30 words.

The class weights are applied to the model's loss functions to tackle class imbalances, and misclassification of underrepresented classes is penalized more severely. The model weighs each class according to its inverse frequency in the training set (Baziotis et al. 2018).

Evaluation metrics Similar to previous emoji prediction studies, this work uses precision (P), recall (R), average F score (F), and accuracy (A) for evaluating the proposed model, which are defined in Eqs. (1, 2, 3, 4) where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$F1 = \frac{2 \times P \times R}{P + R} F1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}. \tag{4}$$

This paper tests the validity of the BARF model by evaluating it against the following models: random forest, logistic regression, stochastic gradient descent (SGD), Naive Bayes, XGBoost, gradient boosting, CNN, LSTM, CNN-LSTM, GRU, BiLSTM, BiLSTM+Attention, and combination of

Table 1 Experimentation of baseline machine learning models on HEP dataset (accuracy in %)

Model	Embeddings				
	FastText	One hot encoding	CharLevel TF-IDF	WordLevel TF-IDF	N-Gram TF-IDF
Random Forest	54.63	53.0	54.21	52.28	41.63
Logistic Regression	45.37	26.59	32.79	33.65	31.43
Support Vector Machine	43.82	9.67	37.64	41.86	35.82
Naive Bayes	33.41	2.78	29.95	30.99	30.67
XGBoost	32.30	32.33	29.39	26.25	28.25
Gradient Boosting	28.25	31.76	30.71	30.82	30.37

Table 2 Comparison of BARF model with baseline deep learning models

Model	Accuracy (%)	Precision	Recall	*F*-score
CNN	30.00	0.15	0.27	0.17
GRU	30.12	0.18	0.28	0.19
LSTM	31.12	0.20	0.31	0.21
CNN LSTM	35.82	0.23	0.37	0.23
BiLSTM	38.80	0.27	0.32	0.27
BiLSTM + Attention	40.10	0.31	0.38	0.30
CNN + BiLSTM +Attention	40.75	0.31	0.38	0.31
CNN + BiGRU + Attention	39.88	0.30	0.37	0.29
BiGRU + Attention + Random Forest	58.85	0.58	0.57	0.57
BARF	**61.14**	**0.66**	**0.59**	**0.59**

Bold indicates the proposed model's name

them (as indicated in Tables 1 and 2). A brief description of these models is as follows:

Random Forest (RF) According to Xu et al. (2012), RF classifiers are well suited for coping with noisy data with a large dimension in text categorization. An RF model comprises a set of decision trees, each trained using a different set of random feature subsets. Given an instance, the RF predicts based on a majority vote of all the trees in the forest.

Logistic Regression(LR) Wright (1995) is a probability-based predictive analytic method. The logistic regression classifier sends the weighted combination of input characteristics via a sigmoid function. The sigmoid function can convert any real number between 0 and 1.

Stochastic Gradient Descent(SGD) classifier is a linear classifier trained with SGD (Kabir et al. 2015). Pure SGD tends to converge to minima with better generalization performances across various NLP problems.

Naive Bayes(NB) classifier is a basic classifier that classifies events based on their probability. It is based on the Bayes theorem, which states that "conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring."

XGBoost is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework (Chen et al. 2015).

Support Vector Machine (SVM) is a supervised machine learning algorithm employed to solve linear and nonlinear problems. SVM separates the data into classes by creating a line of the hyperplane.

GRU (Cho et al. 2014) is a type of RNN network that overcomes the vanishing gradient problem that is present in a regular recurrent neural network by using an update and reset gate.

CNNs recognize patterns across space (Kim 2014). CNNs excel at detecting position-invariant and local patterns. Patterns could be keywords expressing a certain attitude, such as "I hate," or a subject like "Harappa civilization." Thus, CNNs have risen as a paradigm architecture for text classification.

LSTM (Sundermeyer et al. 2012) cell helps in preserving the context and recent dependencies in the text. It can give good results even when the dataset is small.

BiLSTM (Graves and Schmidhuber 2005) is a bidirectional long-term and short-term memory network that is particularly well suited to modeling sequential data. Applying two LSTMs in opposite directions could better understand the context.

BiLSTM+Attention Zhou et al. (2016) utilizes attention mechanism to further enhance BiLSTMs feature learning process. It uses the fact that the model's attention mechanism allows it to focus on and acquire essential information; thus, it delivers good results in text classification.

Experimental results This work experiments with a combination of various machine learning models and embeddings as shown in Table 1 to find the best performing model. It uses embeddings, such as pre-trained Word2Vec, Glove, Word2Vec trained on the HEP dataset, and FastText embeddings. It explores machine learning models, namely random forest, logistic regression, support vector machine, Naive Bayes, XGBoost, and gradient boosting. Table 1 presents the results in terms of accuracy (A), where the random forest algorithm performs significantly better than other machine learning models. Random forest algorithm with FastText embeddings outperforms the Logistic Regression machine learning model by 9%, with the second-best results.

This work also performs experiments using standard deep learning models. The primary concept is to use the most effective machine learning and deep learning models for emoji prediction tasks. A deep learning model understands the text's context, dependencies, and essential sections. The RF machine learning model further classifies the text without overfitting. This work evaluates different deep learning models against four metrics: precision, recall, *F* score (F), accuracy (A) as shown in Table 2. The proposed BiLSTM attention random forest (BARF) model significantly outperforms the baseline models and achieves an accuracy of 61.14%, the precision of 0.66, recall of 0.59, and an F1 score of 0.59. BiLSTM attention random forest (BARF) also outperforms the commonly used BiLSTM attention model by 21% on accuracy, 35% on precision, 21% on recall, and 29% on F1 score.

In the ablation study, each component of the BARF model gives the optimal performance. As shown in Table 2, complete BiLSTM attention random forest (BARF) model performs best among all the models. Performance of the BiLSTM model increases by 2% on accuracy, 4% on precision, 6% on recall, and 3% on F1 score after the addition of attention mechanism, whereas the performance increases by 21% on accuracy, 35% on precision, 21% on recall and 29% on F score for the BiLSTM with Attention layer and RF model.

Hyperparameter optimization To identify optimal hyperparameter values for the BARF's deep learning model, it uses KerasTuner; and employs Scikit-RandomizedSearchCV for random forest. It uses Adam algorithm (Kingma and Ba 2014) as the optimizer and a dropout of 0.2. It adds the BiLSTM in one layer with 512 neurons and a dense layer with 8192 neurons. The loss function is Categorical Cross-Entropy, and the batch size is 32. BARF uses 200 random forest estimators and sets the number of features to 40. The proposed model obtained 61.14% accuracy on the HEP dataset with these hyperparameters.

5 Conclusion and future work

As emojis have become popular and their numbers have increased, there is a huge demand for the emoji prediction model. Several researchers have worked on emoji prediction systems for resource-rich languages such as English. However, there is limited attention for resource-poor and code-mixed languages. This paper offers a new Hinglish Emoji Prediction (HEP) dataset created using Twitter as a corpus and proposes a novel hybrid emoji prediction model for Hinglish, one of the most often used code-mixed languages worldwide. To the best of our knowledge, this is the first attempt to explore emoji prediction for Hinglish. The suggested approach treats the emoji prediction problem as a text classification task.

This study considers the 40 most frequently used emojis in the newly created Hinglish Emoji Prediction (HEP) dataset for the emoji prediction task. The proposed BiLSTM attention random forest (BARF) model uses a BiLSTM network which effectively increases the amount of information available to the network by providing the current and future context and passes this to the attention layer. The attention mechanism increases the weight of words, which helps to improve prediction accuracy. Further, random forest enhances the feature classification. It classifies the features from the attention layer into one of the forty emoji labels. Further, Semantic Similarity and Cosine Similarity Algorithm is used for prediction emoji can be easily identified through the words like 🍰cake, 🍃leaf, 🔥fire. The experimental findings on the HEP dataset show that the proposed BARF model has a better prediction efficacy than the competitive models, and its emoji prediction is more relevant to real life.

Continuing the Hinglish emoji prediction work further, we explored the emoji prediction task as a translation problem instead of a classification problem and used an encoder–decoder-based model to predict multiple emojis for a given plain text input (Himabindu et al. (in press)). In the future, instead of just predicting emoji, we can concentrate on further enhancing the emoji usage experience by introducing sentiment-aware emoji insertion. For given plain text input, the emoji insertion task aims at inserting relevant emojis into the text at proper positions (Lin et al. 2021; Jiang et al. 2020; Kwon et al. 2021).

Declarations

Conflict of interest The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

Aoki S, Uchida O (2011) A method for automatically generating the emotional vectors of emoticons using weblog articles. In: Proceedings 10th WSEAS international conference on applied computer and applied computational science, Stevens Point, Wisconsin, USA, pp 132–136

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

Bali K, Sharma J, Choudhury M, Vyas Y (2014) "I am borrowing ya mixing?" An analysis of english–hindi code mixing in facebook. In: Proceedings of the first workshop on computational approaches to code switching, pp 116–126

Barbieri F, Ballesteros M, Saggion H (2017) Are emojis predictable? arXiv preprint arXiv:1702.07285

Barbieri F, Camacho-Collados J, Ronzano F, Anke LE, Ballesteros M, Basile V, Patti V, Saggion H (2018) SemEval 2018 task 2: multilingual emoji prediction. In: Proceedings of The 12th international workshop on semantic evaluation, pp 24–33

Barbieri F, Espinosa-Anke L, Camacho-Collados J, Schockaert S, Saggion H (2018) Interpretable emoji prediction via label-wise attention lstms. In: Proceedings of the 2018 conference on empirical methods in natural language processing; 2018 Oct 31–Nov 4; Brussels, Belgium. New York: Association for Computational Linguistics; 2018. ACL (Association for Computational Linguistics)

Barbieri F, Espinosa-Anke L, Saggion H (2016) Revealing patterns of twitter emoji usage in Barcelona and Madrid. In: Artificial intelligence research and development, pp 239–244

Barbieri F, Kruszewski G, Ronzano F, Saggion H (2016) How cosmopolitan are emojis? Exploring emojis usage and meaning over different languages with distributional semantics. In: Proceedings of the 24th ACM international conference on multimedia, pp 531–535

Barbieri F, Ronzano F, Saggion H (2016) What does this emoji mean? A vector space skip-gram model for twitter emojis. In: Proceedings of the Tenth international conference on language resources and evaluation (LREC), pp 3967–3972

Barbieri, Francesco and Espinosa-Anke, Luis and Saggion, Horacio (2016) Revealing Patterns of Twitter Emoji Usage in Barcelona and Madrid. Artificial Intelligence Research and Development IOS Press, pp 239-244

Baziotis C, Athanasiou N, Paraskevopoulos G, Ellinas N, Kolovou A, Potamianos A (2018) Ntua-slp at semeval-2018 task 2: predicting emojis using rnns with context-aware attention. arXiv preprint arXiv:1804.06657

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H et al (2015) Xgboost. Extreme gradient boosting. R package version 0.4-2. 1(4):1–4

Choudhary N, Singh R, Bindlish I, Shrivastava M (2018) Contrastive learning of emoji-based representations for resource-poor languages. arXiv preprint arXiv:1804.01855

Choudhary N, Singh R, Rao VA, Shrivastava M (2018) Twitter corpus of resource-scarce languages for sentiment analysis and multilingual emoji prediction. In: Proceedings of the 27th international conference on computational linguistics, pp 1570–1577

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078

Çöltekin Ç, Rama T (2018) Tübingen-oslo at semeval-2018 task 2: Svms perform better than RNNS in emoji prediction. In: Proceedings of the 12th international workshop on semantic Evaluation, pp 34–38

Eisner B, Rocktäschel T, Augenstein I, Bošnjak M, Riedel S (2016) emoji2vec: learning emoji representations from their description. arXiv preprint arXiv:1609.08359

Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm networks. In: Proceedings 2005 IEEE international joint conference on neural networks, 2005., vol 4, pp 2047–2052. IEEE

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm networks. In: Proceedings 2005 IEEE international joint conference on neural networks, 2005., vol 4, pp 2047–2052. IEEE

Guibon G, Ochs M, Bellot P (2018) Emoji recommendation in private instant messages. In: Proceedings of the 33rd Annual Acm symposium on applied computing, pp 1821–1823

Han S, Williamson BD, Fong Y (2021) Improving random forest predictions in small datasets from two-phase sampling designs. BMC Med Inform Decis Mak 21(1):1–9

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Himabindu GSSN, Rao R, Sethia D (2022) Encoder-decoder based multi-label emoji prediction for Code-Mixed Language (Hindi+English). In: 2nd International Conference on Intelligent Technologies (CONIT), pp 1–6. https://doi.org/10.1109/CONIT55038.2022.9848356

Jiang H, Guo A, Ma J (2020) Automatic prediction and insertion of multiple emojis in social media text. In: 2020 International conferences on Internet of Things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData) and IEEE congress on cybermatics (Cybermatics), pp 505–512. IEEE

Kabir F, Siddique S, Kotwal MRA, Huda MN (2015) Bangla text document categorization using stochastic gradient descent (sgd) classifier. In: 2015 international conference on cognitive computing and information processing (CCIP), pp 1–4 . IEEE

Kim Y (2014) Convolutional neural networks for sentence classification. New York University. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980

Kwon J, Kobayashi N, Kamigaito H, Takamura H, Okumura M (2021) Making your tweets more fancy: emoji insertion to texts. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp 770–779

Liebeskind C, Liebeskind S (2019) Emoji prediction for hebrew political domain. In: Companion proceedings of the 2019 world wide web conference, pp 468–477

Lin F, Song Y, Ma X, Min E, Liu B (2021) Sentiment-aware emoji insertion via sequence tagging. IEEE Multimed 28(2):40–48

Ling W, Luís T, Marujo L, Astudillo RF, Amir S, Dyer C, Black AW, Trancoso I (2015)Finding function in form: compositional character models for open vocabulary word representation. arXiv preprint arXiv:1508.02096

Mathur P, Sawhney R, Ayyar M, Shah R (2018) Did you offend me? Classification of offensive tweets in Hinglish language. In: Proceedings of the 2nd workshop on abusive language online (ALW2). Association for Computational Linguistics, Brussels, Belgium

Parshad RD, Bhowmick S, Chand V, Kumari N, Sinha N (2016) What is India speaking? Exploring the “hinglish'' invasion. Phys A Statist Mech Appl 449:375–389

Peng D, Zhao H (2021) Seq2emoji: a hybrid sequence generation model for short text emoji prediction. Knowl-Based Syst 214:106727

Pohl H, Domin C, Rohs M (2017) Beyond just text: semantic emoji similarity modeling to support expressive communication. ACM Trans Comput-Human Inter (TOCHI) 24(1):1–42

Qi Y (2012) Random forest for bioinformatics. In: Ensemble Machine Learning Springer Science & Business Media, pp 307

Ronzano F, Barbieri F, Wahyu Pamungkas E, Patti V, Chiusaroli F, et al (2018) Overview of the evalita 2018 Italian emoji prediction (itamoji) task. In: 6th evaluation campaign of natural language processing and speech tools for Italian. Final Workshop, EVALITA 2018, vol 2263, pp 1–9 . CEUR-WS

Sundermeyer M, Schlüter R, Ney H (2012) Lstm neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association

Tomihira T, Otsuka A, Yamashita A, Satoh T (2018) What does your tweet emotion mean? Neural emoji prediction for sentiment analysis. In: Proceedings of the 20th international conference on information integration and web-based applications & services, pp 289–296

Tomihira T, Otsuka A, Yamashita A, Satoh T (2020) Multilingual emoji prediction using BERT for sentiment analysis. International Journal of Web Information Systems Emerald Publishing Limited

Vidal L, Ares G, Jaeger SR (2016) Use of emoticon and emoji in tweets for food-related emotional expression. Food Qual Prefer 49:119–128

Vijay D, Bohra A, Singh V, Akhtar SS, Shrivastava M (2018) Corpus creation and emotion prediction for hindi–english code-mixed social media text. In: Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: student research workshop, pp 128–135

Wijeratne S, Balasuriya L, Sheth A, Doran D (2017) A semantics-based measure of emoji similarity. In: Proceedings of the international conference on web intelligence, pp 646–653

Wright RE (1995) Logistic regression reading and understanding multivariate statistics American Psychological Association, pp 217–244

Wu C, Wu F, Wu S, Huang Y, Xie X (2018) Tweet emoji prediction using hierarchical model with attention. In: Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers, pp 1337–1344

Xie R, Liu Z, Yan R, Sun M (2016) Neural emoji recommendation in dialogue systems. arXiv preprint arXiv:1612.04609

Xu B, Guo X, Ye Y, Cheng J (2012) An improved random forest classifier for text categorization. J Comput 7(12):2913–2920

Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1480–1489

Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short Papers), pp 207–212

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.