Building and Assessing Emoji Prediction Models for English and Spanish Tweets

1 Introduction

An emoji is a small digital image or icon that represents an idea, emotion, object, or concept in electronic communication. Originally developed in Japan, the term "emoji" combines the Japanese words "e" (meaning picture) and "moji" (meaning character) [1]. As pictorial symbols, emojis cover a broad spectrum of subjects, from facial expressions to objects, characters, activities, and beyond. Their visual nature enhances online communication, making interactions more vibrant and expressive. As for real-life application examples, businesses, recognizing the power of visual messages in advertisement, actively incorporate them to boost engagement on platforms like Twitter, effectively adding a visually appealing and relatable layer to their online communication strategies. More importantly, emojis effectively convey a wide range of emotions, enhancing textual communication by visually expressing sentiments and feelings. Because they can provide valuable cues for understanding and interpreting the emotional tone of written text, emojis become crucial for sentiment analysis in Natural Language Processing (NLP). As the scientific community continues to explore and refine sentiment analysis techniques, the integration of emojis offers valuable hints and insights for capturing nuanced emotions in text data. In 2018, Barbieri et al. [2] proposed the shared task of emoji prediction and analyzed nearly 50 model submissions. The study highlighted the competitive nature of the FastText baseline model and recommended future research to explore additional languages. Followingly, Lou et al. [3] investigated sentiment analysis with emojis, proposing a model based on Bidirectional Long Short-Term Memory (BiLSTM) that eliminates feature engineering and captures emojis' impact on text sentiment using an attention approach. Recently, Zhang et al. [4] introduced TwHIN-BERT, a multilingual language model tailored for social media trained on 7 billion tweets across diverse languages. The model, incorporating a social objective from Twitter's heterogeneous information network, offers a unique approach to handling short, noisy user-generated text.

Recognizing the importance of precise emoji prediction in diverse linguistic contexts, my research project poses the following question: how can emoji prediction models be effectively developed and refined to accommodate the linguistic variations in English and Spanish tweets? To answer the question, the project implements models that take one tweet sentence (such as "I love pizza!") as input and find the most appropriate emoji to be used with this sentence (🍕), starting with a baseline model using glove word embeddings and then trying to improve the performance with LSTM. This task is motivated by the understanding that accurate and culturally sensitive emoji predictions are crucial for developing efficient recommendations, aiming to facilitate smoother online conversations and enhance user experiences.

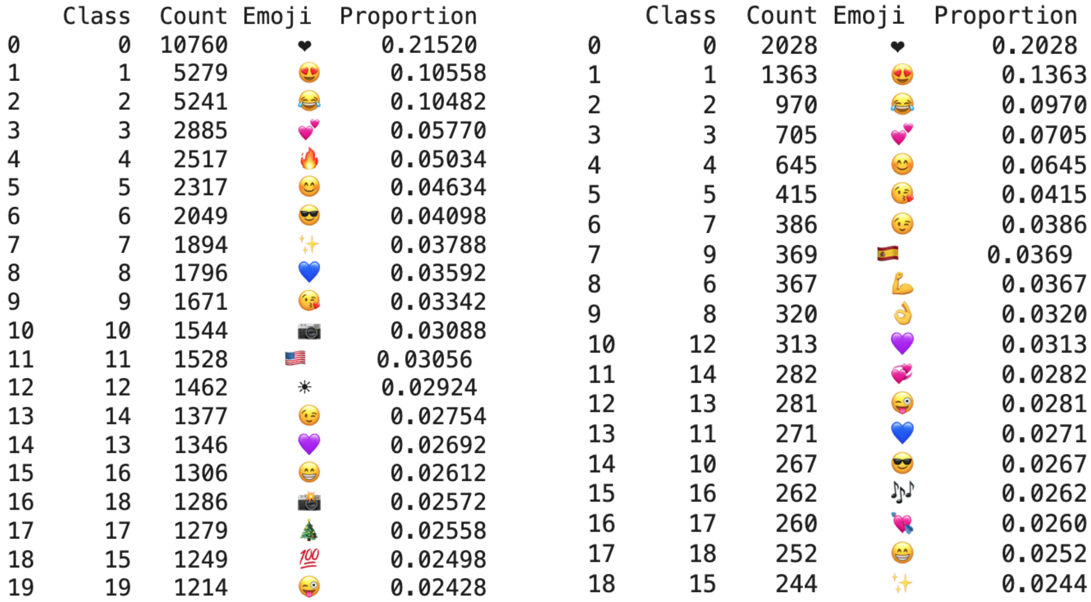
English and Spanish were chosen for examination as they are the two most commonly used languages in the United States. The significance of this project lies in its potential to foster the development of effective emoji recommendations. By achieving this, the project seeks to contribute to facilitating seamless online conversations and enhancing user experiences with emoji recommendation systems.

2 Data Preparation

The data was collected from the International Workshop on Semantic Evaluation (SemEval) 2018 Task 2, Multilingual Emoji Prediction. [5] The dataset, comprising tweets geolocalized within the United States and Spain from October 2015 to February 2017, aligns temporally and geographically with Twitter usage in these regions. The credibility and suitability of SemEvalTask for predicting emojis in both English and Spanish are established in the reference paper 'SemEval-2018 Task 2: Multilingual Emoji Prediction.' This paper, authored by reputable researchers and published in the Proceedings of the 12th International Workshop on Semantic Evaluation, adds a robust academic underpinning to our project.

The project primarily utilized data from two folders provided in SemEval Task 2: "trial" and "mapping." There were four files within the original trial folder, two for English and two for Spanish. One of the two files for each language contains tweets on each row and the other file associates a corresponding emoji with each line of tweet. The files containing tweets are renamed as “us_input.csv” (“es_input.csv” if it is in Spanish) and the files with emojis are renamed with “us_label.csv” (“es_label.csv” if it is in Spanish). Each English file contains 50000 tweets and each Spanish file contains 10000 tweets. Each tweet in the dataset contains one and only one emoji of the 20 most frequent emojis in either language. To better organize and make plots of the data using pandas DataFrame, I wrote a function **remove_specials** to transform the files from text to CSV formats.

To make the task simpler with numbers as class labels, the mapping folder contains files that map emojis to their index class and word descriptions for each language. The index class of each word is their rank of usage frequency in either language. The top 20 emojis are different across the English and Spanish corpora as expected because emoji usage varies based on cultural and linguistic factors as shown below.



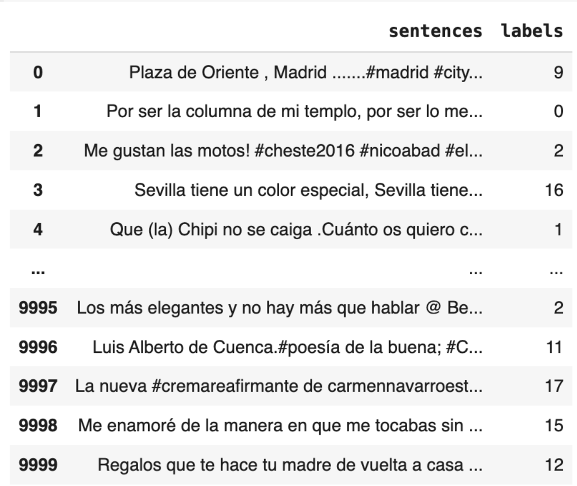
```
    Class  Count Emoji  Proportion
0       0  10760    ❤      0.21520
1       1   5279    😍      0.10558
2       2   5241    😂      0.10482
3       3   2885    💕      0.05770
4       4   2517    🔥      0.05034
5       5   2317    😊      0.04634
6       6   2049    😎      0.04098
7       7   1894    ✨      0.03788
8       8   1796    💙      0.03592
9       9   1671    😘      0.03342
10     10   1544    📷      0.03088
11     11   1528   🇺🇸     0.03056
12     12   1462    ☀      0.02924
13     14   1377    😉      0.02754
14     13   1346    💜      0.02692
15     16   1306    😁      0.02612
16     18   1286    📸      0.02572
17     17   1279    🎄      0.02558
18     15   1249    💯      0.02498
19     19   1214    😜      0.02428
```

```
    Class  Count Emoji  Proportion
0       0   2028    ❤      0.2028
1       1   1363    😍      0.1363
2       2    970    😂      0.0970
3       3    705    💕      0.0705
4       4    645    😊      0.0645
5       5    415    😘      0.0415
6       7    386    😉      0.0386
7       9    369   🇪🇸     0.0369
8       6    367    💪      0.0367
9       8    320    👌      0.0320
10     12    313    💜      0.0313
11     14    282    💞      0.0282
12     13    281    😜      0.0281
13     11    271    💙      0.0271
14     10    267    😎      0.0267
15     16    262    🎶      0.0262
16     17    260    💘      0.0260
17     18    252    😁      0.0252
18     15    244    ✨      0.0244
```

Figure 1. Distribution of the top 20 emojis in English (left) and Spanish (right)

Users from distinct language backgrounds may prefer emojis that align with their cultural context, leading to unique patterns in emoji selection. As an illustrative instance, the Spanish flag emerges as one of the prominent emojis within the top 20 selections in the context of Spanish communication but it is not one of the top 20 emojis in English. This discrepancy underscores the cultural differences that influence users' emoji preferences and usage patterns.

Due to a data processing issue not specified by the organizer of the task, we only consider 19 emojis in the Spanish task (from 0 to 18 where the top emoji is omitted). Below is an overview of how each tweet in the Spanish dataset corresponds to an index class of an emoji.



	sentences	labels
0	Plaza de Oriente , Madrid ....#madrid #city...	9
1	Por ser la columna de mi templo, por ser lo me...	0
2	Me gustan las motos! #cheste2016 #nicoabad #el...	2
3	Sevilla tiene un color especial, Sevilla tiene...	16
4	Que (la) Chipi no se caiga .Cuánto os quiero c...	1
...	...	...
9995	Los más elegantes y no hay más que hablar @ Be...	2
9996	Luis Alberto de Cuenca.#poesía de la buena; #C...	11
9997	La nueva #cremareafirmante de carmennavarroest...	17
9998	Me enamoré de la manera en que me tocabas sin ...	15
9999	Regalos que te hace tu madre de vuelta a casa ...	12

Figure 2. Each Spanish tweet corresponding to one emoji class (label)

Additionally, a "glove.6B.50d.txt" file from Kaggle [6] is included for word embedding purposes. The file provides pre-trained GloVe (Global Vectors for Word Representation) embeddings with a dimensionality of 50, which allows for a more nuanced and context-aware representation of words in the embedding space. It leverages the rich semantic information that can enhance the performance of models in capturing subtle linguistic nuances across both English and Spanish text data.

3 Model

Baseline model

Because the emoji task is essentially predicting the most appropriate class among 20 classes, it constitutes a multiclass classification problem. A simple neural network, equipped with a straightforward architecture and softmax activation, is well-aligned with the nature of multiclass classification tasks. The implementation consists of a training function (model) and a prediction function (predict). The training function initializes weights and biases and iteratively updates parameters using gradient descent. The prediction function computes predictions based on learned parameters. The model uses an average sentence representation and softmax activation for predictions. Average sentence representation is a choice of simplicity and computational efficiency since we are building a baseline model. Its adaptability to variable-length tweets is especially suitable for our task as emoji prediction resembles text classification. We used softmax activation in the end because our task is a multiclass classification and the model's raw output needs to be converted into a probability distribution over multiple classes for final prediction.

Below is an illustration of the baseline model structure. The input of the model is a tweet without an emoji with a max length constraint of 100 (according to our dataset, the length of any tweet will be no longer than 100). During the training process, the model first converts the words in the input sentence into separate Glove representations. It then averages the word vector representations into a single vector to represent the sentence meaning. Afterwards, the single vector will be passed through the forward propagation and the model will compute the loss and backpropagate to update the softmax's parameters. The cross-entropy loss is computed by the formula below:

$$\mathcal{L}^{(i)} = -\sum_{k=0}^{n_y-1} Y_{oh,k}^{(i)} * log(a_k^{(i)})$$

Figure 3. Multi-class Cross-Entropy Loss [7]

The output is a probability vector of shape (1,20) and with the argmax layer, we can then determine which index class, or which emoji, is most likely to follow the tweet. After conducting multiple experiments, I discovered that setting the learning rate to 0.01 and the number of iterations to 1000 yields comparatively good results.

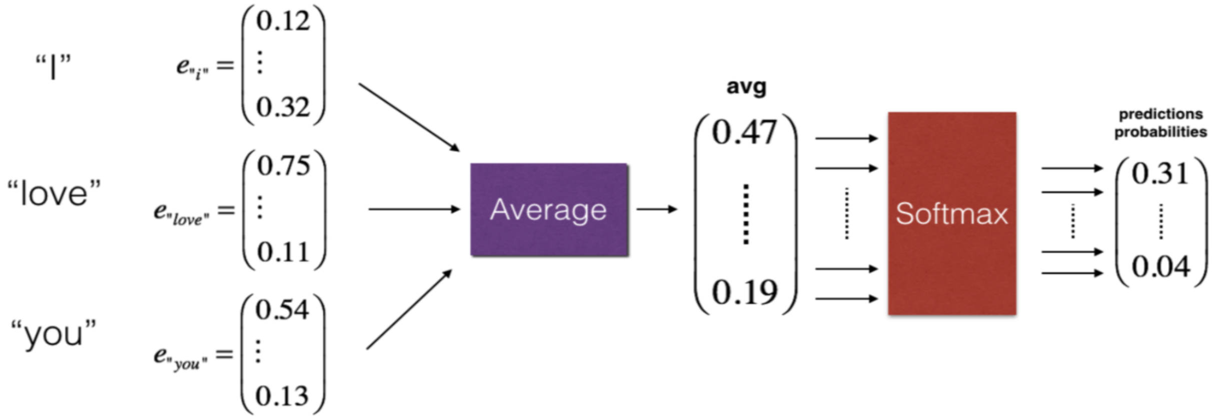


Figure 4. Baseline model structure

LSTM model

The attempt to improve the baseline model employing a recurrent neural network (RNN) with Long Short-Term Memory (LSTM) layers for emoji prediction is motivated by its demonstrated efficacy in capturing the nuanced semantics of emojis, as substantiated by the findings of the referenced work. According to Barbieri et al. [8], LSTMs outperform not only a baseline model but also humans in the task of predicting emojis associated with text-based tweet messages. In alignment with this foundation, a 2-layer LSTM sequence classifier was built for this project. The architecture comprises two Long Short-Term Memory (LSTM) layers with 128 units each, configured to capture sequential dependencies in the input data. Dropout layers with a rate of 0.5 follow both LSTM layers, serving as regularization mechanisms to mitigate overfitting. The model includes a dense layer with 20 units, aligning with the number of target emojis, and applies the softmax activation function to output probabilities for each emoji class. Further refinement may involve adjusting the dropout rate to enhance its predictive performance for emoji classification. Similarly, the input of the model is a tweet without an emoji, no longer than a length of 100 and the output will be a probability distribution over the 20 target emojis.

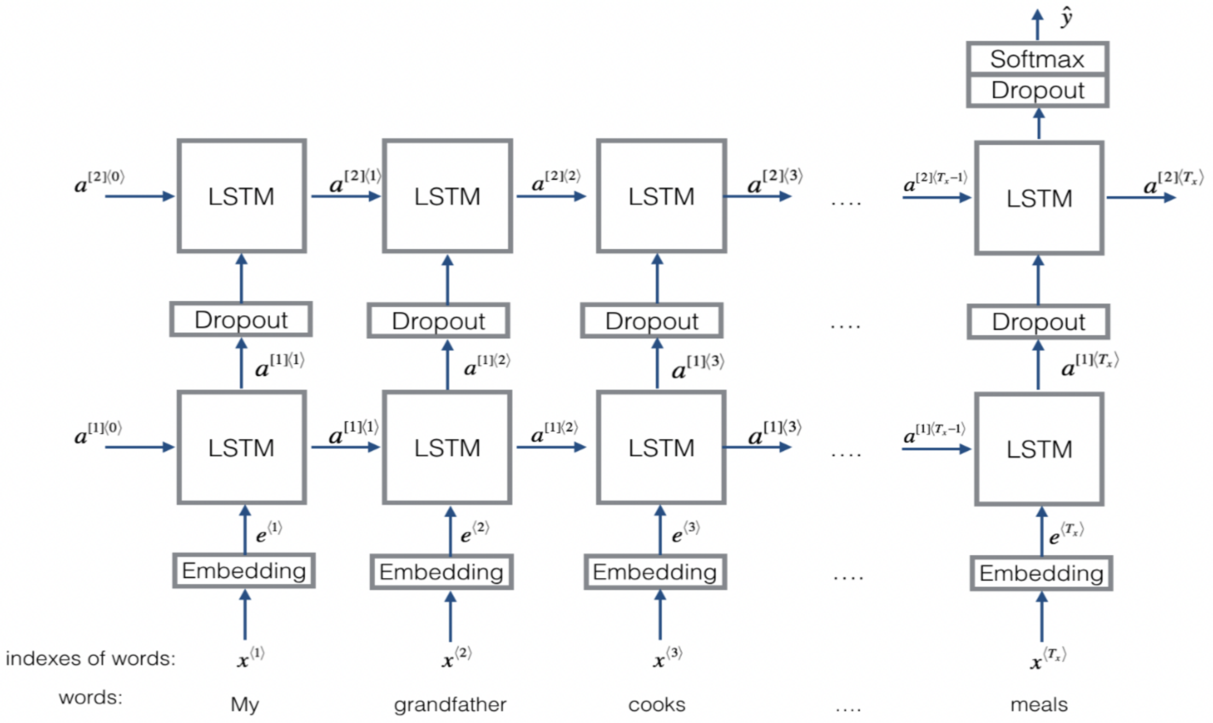

$\hat{y}$
Softmax
Dropout
$a^{[2]\langle 0\rangle}$
LSTM
$a^{[2]\langle 1\rangle}$
LSTM
$a^{[2]\langle 2\rangle}$
LSTM
$a^{[2]\langle 3\rangle}$
....
$a^{[2]\langle T_x-1\rangle}$
LSTM
$a^{[2]\langle T_x\rangle}$
Dropout
Dropout
Dropout
....
Dropout
$a^{[1]\langle 1\rangle}$
$a^{[1]\langle 2\rangle}$
$a^{[1]\langle 3\rangle}$
$a^{[1]\langle T_x\rangle}$
$a^{[1]\langle 0\rangle}$
LSTM
$a^{[1]\langle 1\rangle}$
LSTM
$a^{[1]\langle 2\rangle}$
LSTM
$a^{[1]\langle 3\rangle}$
....
$a^{[1]\langle T_x-1\rangle}$
LSTM
$e^{\langle 1\rangle}$
$e^{\langle 2\rangle}$
$e^{\langle 3\rangle}$
$e^{\langle T_x\rangle}$
Embedding
Embedding
Embedding
....
Embedding
indexes of words:
$x^{\langle 1\rangle}$
$x^{\langle 2\rangle}$
$x^{\langle 3\rangle}$
$x^{\langle T_x\rangle}$
words:
My
grandfather
cooks
....
meals

Figure 5. 2-layer LSTM sequence classifier [9]

The embedding layer in this model is initialized also using pre-trained GloVe 50-dimensional vectors. The vocabulary size is determined by the length of the word-to-index mapping plus one, adhering to the requirement of the Keras embedding layer. The embedding dimension is specified based on the shape of the pre-trained GloVe vectors. An embedding matrix is created as an array of zeros and is populated by assigning the GloVe word vectors to the matrix based on the vocabulary. The Keras Embedding layer is then defined with the specified vocabulary size and embedding dimension and set as non-trainable to retain the pre-trained vectors.

4 Metrics

In addressing the single-label classification problem of emoji prediction, a comparative analysis of accuracy, precision, recall, and F-score was conducted to assess the model. To ensure impartiality across all emoji classes, the macro-average approach is employed, treating each class equitably to mitigate potential biases related to class frequency disparities. This methodology ensures a fair and comprehensive evaluation of the model's efficacy in predicting the entire spectrum of 20 emojis.

Accuracy serves as a holistic metric, quantifying the model's overall correctness by dividing the count of accurately predicted instances by the total dataset size, expressed as a percentage. Precision provides insights into the model's accuracy in predicting a specific emoji, calculated by dividing the number of correctly predicted instances by the total instances predicted as that emoji, thereby assessing its capacity to minimize false positives. Conversely, recall measures the model's proficiency in capturing instances of a specific emoji, computed as the ratio of correctly predicted instances to the total instances of that emoji in the dataset. Additionally, we computed the F1 score, a metric that strikes a balance between precision and recall. It is calculated as follows:

$$F = \frac{2Precision \cdot Recall}{Precision + Recall}$$

Figure 6. F-score formula [10]

In cases where achieving a balance between precision and recall is crucial, the F1 score provides a singular measure that captures both aspects of the model's performance. More importantly, given the skewed distribution of data in both English and Spanish, the F1 score emerges as a more insightful metric. It provides a nuanced understanding of results, particularly for the low-frequency classes, as it inherently rewards systems capable of accurately predicting instances within these less frequent categories.

The baseline model's success criteria are established considering the probability of randomly guessing an emoji within a set of emojis. For the English task, an accuracy exceeding 5% is deemed successful, aligning with the probability of 1 out of 20. Literature [2] reports that baseline model accuracies range from 3.92% to 42.18% so any accuracy within this interval for English test data would be considered a satisfactory performance. Similarly, the success baseline for the model on Spanish test data would be higher than randomly guessing portability, which is roughly 5%. The literature-reported accuracies range from 13.81% to 34.85%, so any accuracy within this interval for Spanish test data would be considered a satisfactory performance. Extra expectations for the LSTM model involve surpassing the baseline's performances in both English and Spanish tasks.

5 Results

The baseline model performed exceptionally well on a tiny dataset with 20 English tweets, reaching an accuracy of 100% in the end. Across 900 epochs, the neural network exhibited substantial improvement. Initiating with a loss of 2.73 in the first epoch, a notable reduction occurred, reaching 0.35 with a corresponding accuracy of 95% by the 100th epoch. The model continued to refine, achieving a minimal loss of 0.04 and 100% accuracy by the 900th epoch. This underscores the model's convergence and exceptional predictive precision for the 20 tweets, illustrating its nuanced understanding of the small input data.

However, challenges emerged when training with real-world data we have. Despite augmenting the number of epochs, the loss failed to exhibit the expected decrease and, rather concerning, displayed an increasing trend from 3.36 to 3.42 while the accuracy also increased from 23.66% to 24.63% for English tweets. This anomaly prompted an investigation into the high learning rate's influence. Attempts were made to mitigate the issue by reducing the learning rate from 0.01 to 0.0001 and exploring intermediary values. Surprisingly, the loss persisted in its upward trend, suggesting that the model's simplicity might be a limiting factor in capturing the intricacies of the more challenging dataset. The fluctuations in both loss and accuracy could also arise from the inherent stochastic nature of the training process in deep learning, wherein randomness in data sampling and other factors contribute to variations in training. Comparatively, the issue is more severe with the Spanishdta possibly because of its less available training data. The training accuracy of Spanish data increased from 0.21 to 0.23, with a bad increase in loss from 3 to almost 6.

When the model is applied to the test data, it gains a decent accuracy score of around 20% on the task of predicting from 20 classes in both English and Spanish, compared to all baseline model performance in the SemEval task2 [2] ranging from 3.92% to 42.18%. For the evaluation metrics in the context of the English task, the precision, measuring the accuracy of the model's positive predictions, stands at 0.17. This implies that approximately 17% of the predicted emojis were indeed correct, showcasing the model's ability to make accurate selections. However, the recall, capturing the model's capacity to identify all relevant instances, is 0.09. This indicates that the model captures only about 9% of the total instances of the emojis present in the test set. Consequently, the F1 score, a metric that balances precision and recall, is computed at 0.076, suggesting the relatively poor performance of the simple model to the emoji task.

Upon a deeper examination of the confusion matrix, a consistent distribution pattern is found in English and Spanish datasets. In the English confusion matrix, the predominant predictions made by the model correspond to class 0. This tendency raises concerns about potential imbalances in the training data, evident in a right-skewed distribution. Specifically, the proportion of instances belonging to class 0 is markedly imbalanced, likely influencing the model's inclination to predict this class.

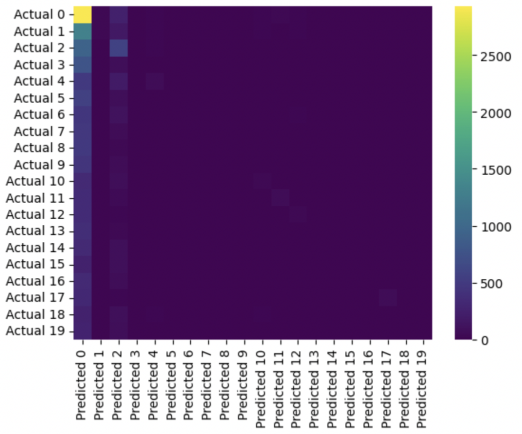


Figure 7. Confusion matrix for English tweets

For LSTM data, the model increases from 0.21 to 0.23 with a steady decrease in loss from 2.75 to 2.73. Although the accuracy is slightly slower than that obtained with the baseline mode, it demonstrates more robust performance in capturing temporal dependencies and handling sequential data. The steady decrease in loss indicates that the LSTM is effectively learning and adapting to the sequential nature of the data. In scenarios where preserving the temporal relationships within the data is paramount as in the emoji prediction task, sacrificing a marginal amount of accuracy for improved sequence modeling can be a worthwhile trade-off. Eventually, the model achieved an accuracy of 0.22 on the test data.

The weighted average precision of 0.05 in the classification report for English tweets indicates that, on average, only 5% of the model's positive predictions are accurate. With a recall of 0.22, the model captures approximately 22% of the total instances of emojis present in the test set. The resulting F1 score of 0.08 suggests a suboptimal balance between precision and recall, indicative of the model's challenges in accurately predicting emojis. Overall, these metrics collectively highlight the limitations in the model's performance for emoji prediction in English tweets, indicating the need for further investigation and potential model refinement to improve its overall predictive capabilities.

The accuracy on the Spniash test set is 19%, slightly worse than the result of English tweets. The loss in training with Spanish tweets also steadily decreased while the accuracy score stabilized around 0.21. The weighted average precision of 0.04 in the classification report for Spanish tweets suggests that, on average, only 4% of the model's positive predictions are accurate. With a recall of 0.19, the model captures approximately 19% of the total instances of emojis present in the test set. The resulting F1 score of 0.06 indicates a suboptimal balance between precision and recall, reflecting challenges in accurately predicting emojis in Spanish tweets.

6 Conclusion

In conclusion, the baseline model exhibited remarkable performance when trained on a small dataset of 20 English tweets, achieving 100% accuracy and showcasing a nuanced understanding of the input data through substantial improvements across 900 epochs. However, when faced with real-world data, challenges emerged, as the model struggled to adapt to the complexities of a larger and more diverse dataset. Despite attempts to address issues such as an increasing loss trend and fluctuations in accuracy by adjusting the learning rate, the simplicity of the model and the stochastic nature of deep learning contributed to limitations in capturing the intricacies of the more challenging dataset.

This challenge became more pronounced with the Spanish dataset, marked by a smaller size of training data, leading to increased loss and lower accuracy. Evaluation metrics of precision, recall, and F1 scores in the context of both languages reveal the model's struggles in making accurate predictions and capturing relevant instances.

The introduction of LSTM data demonstrated promising results, showcasing a steady decrease in loss and emphasizing the model's robustness in capturing temporal dependencies and handling sequential data. While the accuracy was slightly slower than the baseline model, the LSTM's superior performance in sequence modeling proved advantageous for tasks prioritizing temporal relationships, such as emoji prediction.

Several potential threats to the validity of the findings in this project should be acknowledged. Sampling bias could be present if the dataset used is not representative of the broader population of tweets. The imbalanced distribution of classes, particularly the overrepresentation of class 0 in both datasets, may influence the model's learning process and impact observed accuracy, potentially leading to skewed results. The small size of the dataset, especially for Spanish tweets, raises concerns about the generalizability of the findings, as models trained on smaller datasets might struggle to capture the complexity of language patterns. Sensitivity to hyperparameters, such as learning rates, was explored but not comprehensively optimized. As for future development, utilizing more extensive and diverse datasets that accurately represent linguistic variability and temporal trends, along with addressing class imbalances, will contribute to a more robust understanding of model performance.

In summary, the study underscores the importance of considering dataset characteristics, model complexity, and the nature of the task at hand in designing effective models for emoji prediction, especially when confronted with real-world data variations and linguistic challenges. Further refinements in model architecture and training strategies may be necessary to address the observed limitations and enhance predictive capabilities, particularly in diverse linguistic contexts.

References

[1] Qiyu Bai, Qi Dan, Zhe Mu, and Maokun Yang. 2019. A systematic review of emoji: Current research and future perspectives. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02221/full?fbclid=IwAR389o07jrYF-GGEM6hibVJ3hR1cZnqiIQa0mG8re2Aq5JIwg3H33VBO4Ho

[2] Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval 2018 Task 2: Multilingual Emoji Prediction. In Proceedings of the 12th International Workshop on Semantic Evaluation, pages 24–33, New Orleans, Louisiana. Association for Computational Linguistics. https://aclanthology.org/S18-1003

[3] Yinxia Lou, Yue Zhang, Fei Li, Tao Qian, and Donghong Ji. 2020. Emoji-Based Sentiment Analysis Using Attention Networks. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 19, 5, Article 64 (September 2020), 13 pages. https://doi.org/10.1145/3389035

[4] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations at Twitter. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23). Association for Computing Machinery, New York, NY, USA, 5597–5607. https://doi.org/10.1145/3580305.3599921

[5] Anon. 2018. Competition. (2018). Retrieved December 19, 2023 from https://competitions.codalab.org/competitions/17344

[6] Ashish Lal. 2018. Glove.6b.50d.TXT. (January 2018). Retrieved December 19, 2023 from https://www.kaggle.com/datasets/watts2/glove6b50dtxt/

[7] Iffat Zafar, Giounona Tzanidou, Richard Burton, Nimesh Patel, and Leonardo Araujo. Hands-on convolutional neural networks with TensorFlow. https://www.oreilly.com/library/view/hands-on-convolutional-neural/9781789130331/7f34b72e-f571-49d2-a37a-4ed6f8011c93.xhtml

[8] Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are Emojis Predictable?. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 105–111, Valencia, Spain. Association for Computational Linguistics.
https://aclanthology.org/E17-2017/

[9] Ronrest. 2017. LSTM dropout - clarification of last layer. (July 2017). Retrieved December 20, 2023 from https://discuss.pytorch.org/t/lstm-dropout-clarification-of-last-layer/5588

[10] Nan Wang, Jin Wang, and Xuejie Zhang. 2018. Ynu-HPCC at Semeval-2018 task 2: Multi-ensemble BI-GRU model with attention mechanism for multilingual emoji prediction. Proceedings of The 12th International Workshop on Semantic Evaluation (2018). DOI:http://dx.doi.org/10.18653/v1/s18-1073