G-LEAP Journal Experiments Report

2022.1

Jianfeng Hou

houjf@shanghaitech.edu.cn

1 Emoji Prediction Datasets

We train the emoji prediction models using different datasets (at least not one single dataset) for the following two reasons:

1. In real-world scenarios, the models have already been pre-trained when deploying our edge-assisted emoji prediction system. The emoji prediction models may be trained by different groups of people using different devices (cloud servers, edge servers, or even end-devices). The collected dataset for training one emoji prediction model probably differ from others.
2. Using different training datasets can increase the diversity of our emoji prediction models.

We use the following emoji prediction datasets:

1. Celebrity Profiling (Emoji Prediction: Extensions and Benchmarking)
2. Twitter (SemEval-2018 Task 2)
3. Twitter (Twitter Emoji Prediction | Kaggle)
4. Twitter (DeepMoji), **TODO**

2 Emoji Prediction Models

The emoji prediction models can differ in the following three dimensions:

1. Base Model;
2. Vectorization Method;

3. Training Dataset.

Thus, the trained emoji prediction models used in our experiments can be listed clearly in Table 1.

Table 1: The trained emoji prediction models used in our experiments.

Index	**Base Model**	**Validation Accuracy**	**Model Size**	**File**
0	SVM	10462/50000 = 20.92%	768K	`statistical/svm.pkl`
1	Naïve Bayes	11110/50000 = 22.22%	1.3M	`statistical/naive_bayes.pkl`
2	Decision Tree	6568/50000 = 13.14%	1.1M	`statistical/decision_tree.pkl`
3	RNN-1	12781/50000 = 25.56%		`neural/rnn/rnn_1.pkl`
4	RNN-2	13265/50000 = 26.53%		`neural/rnn/rnn_2.pkl`
5	LSTM-1	14255/50000 = 28.51%		`neural/lstm/lstm_1.pkl`
6	LSTM-2	14677/50000 = 29.35%		`neural/lstm/lstm_2.pkl`
7	Bi-LSTM	15342/50000 = 30.68%		`neural/lstm/bi_lstm.pkl`
8	BERT	16532/50000 = 33.06%	518M	`neural/transformer/bert.pkl`
9	RoBERTa	16441/50000 = 32.88%	477M	`neural/transformer/roberta.pkl`

2.1 Statistical Models

Currently we only use Bag of Words (BoW) as the vectorization method for training all the statistical models.

Table 2: The trained emoji prediction models used in our experiments.

Index	Base Model	Vectorization Method	Training Dataset	Model Size	Test Accuracy	Inference Time
0	Naïve Bayes					
1	SVM					
2	BERT					
3	RoBERTa					
4	MLP					
5	LSTM					
6	Bi-LSTM					
7	RNN					
8	DeepMoji					
9	fastText					

2.1.1 Naïve Bayes

2.1.2 Support Vector Machine (SVM)

2.1.3 Decision Tree

2.2 Neural Models

In our experiments, we use the following vectorization method for training MLP models:

- Word embeddings: GloVe embeddings.
- Sequence embeddings: mean of all words in a sentence. Reference

2.2.1 Multi-Layer Perceptron (MLP)

2.2.2 Recurrent Neural Network (RNN)

2.2.3 Long-Short Term Memory (LSTM)

2.2.4 Transformers

2.3 Vectorization

1. Bag of Words / Characters
2. One-hot encoding
3. word embeddings + sequence embeddings

3 Miscellaneous

3.1 Experiments Observations

Neural models with smaller batch size can converge quicker, and produce higher accuracy.