Analyzing and Predicting Emoji Usages in Social Media

Peijun Zhao
zhaopeijun0328@163.com
Department of Computer Science and Technology, Tsinghua University
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)

Jia Jia*
jjia@tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)

Yongsheng An
316991611@qq.com
Academy of Arts & Design, Tsinghua University

Jie Liang
liang-j17@tsinghua.edu.cn
Academy of Arts & Design, Tsinghua University

Lexing Xie
lexing.xie@anu.edu.au
Department of Computer Science, Australian National University

Jiebo Luo
jiebo.luo@gmail.com
Department of Computer Science, University of Rochester

ABSTRACT

Emojis can be regarded as a language for graphical expression of emotions, and have been widely used in social media. They can express more delicate feelings beyond textual information and improve the effectiveness of computer-mediated communication. Recent advances in machine learning make it possible to automatic compose text messages with emojis. However, the usages of emojis can be complicated and subtle so that analyzing and predicting emojis is a challenging problem. In this paper, we first construct a benchmark dataset of emojis with tweets and systematically investigate emoji usages in terms of tweet content, tweet structure and user demographics. Inspired by the investigation results, we further propose a multitask multimodality gated recurrent unit (mmGRU) model to predict the categories and positions of emojis. The model leverages not only multimodality information such as text, image and user demographics, but also the strong correlations between emoji categories and their positions. Our experimental results show that the proposed method can significantly improve the accuracy for predicting emojis for tweets (+9.0% in F1-value for category and +4.6% in F1-value for position). Based on the experimental results, we further conduct a series of case studies to unveil how emojis are used in social media.

KEYWORDS

Emoji; GRU; Multimodality; Multitask

ACM Reference Format:
Peijun Zhao, Jia Jia, Yongsheng An, Jie Liang, Lexing Xie, and Jiebo Luo. 2018. Analyzing and Predicting Emoji Usages in Social Media. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3186344

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186344

1 INTRODUCTION

Emojis have been widely used in social networks to express more delicate feelings beyond textual information and make computer-mediated communication more effective. Vulture's Lindsey Weber, who co-curated the "Emoji" art show[1], says that she use emojis in personal emails all the time, because she feel like she is softening the email. According to the study in Instagram[4], emojis are present in up to 40% of online messages in many countries. For example, "Face with Tears of Joy", an emoji that means the person has an extremely good mood, is regarded as the 2015 word of the year by The Oxford Dictionary.

Since most people have experiences in using emojis for posting tweets, the problem of automatically generating emojis becomes interesting and useful for online users. Existing works usually study emoji usages based on textual information. [8] uses Affective Trajectory Model for emoji category recommendation based on textual information. [16] analyzes the emotion of the textual information published by users and proposes an emoji recommendation method based on the emotive statements of users and their past selections of emojis. However, whether there are regularities or norms in the choices of emojis and what information is related to them are still open problems, which make the problem of generating emojis challenging.

The usages of emojis depend on many complex factors. Since emojis are an integral part of tweets and play an important role in expressing emotion, the choices of emojis are mainly affected by the following factors: 1) Tweet content. Not only textual information, but also visual information can enrich the expression of emotion for tweets. How can the multimodality information affect the choices of emojis? 2) Tweet structure. Different emojis may appear at different positions in tweets. Whether there are correlations between the categories and positions of emojis has not yet been confirmed. 3) User demographics. Different users may have different habits on using emojis. How can user demographics such as geographica region information influence assigning emojis?

*Corresponding author
[1] http://www.emojishow.com/

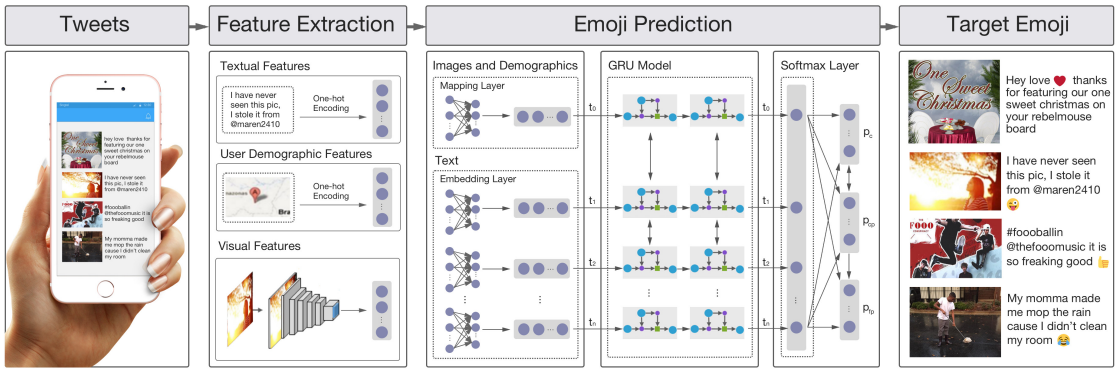


Figure 1: Overview of the Proposed Emoji Prediction Method. Far-left: Example tweets from Twitter. Mid-left: Textual, visual and user demographic features. Mid-right: The mmGRU model. Far-right: Predictions results.

In this paper, we first construct a benchmark dataset of emojis from Twitter. It contains 164,880 tweets of the 35 most frequently used emojis. Using the dataset, we systematically study emoji usages with tweet content, tweet structure and user demographics and obtain a series of data-driven observations. Inspired by the observations, we propose a multitask multimodality gated recurrent unit (mmGRU) model (Figure 1) to quantitatively study the correlations between emojis and the above three aspects. The multimodality function is used to incorporate multimodality information such as text, image, and user demographics, while the multitask framework contributes to leveraging the strong correlations between emoji categories and their positions for improving the performance of both tasks. We apply the proposed model to predict the emoji categories and positions of online tweets in our dataset. Our experimental results show that our model can significantly improve the prediction performance of emoji category by +9.0% in F1-value and emoji position by +4.6% in F1-value, on the average, compared with several alternative baselines. Based on the experimental results, we further conduct a series of case studies to reveal how emojis are used in social media.

Our main contributions are as follows:

- We construct a large-scale benchmark dataset for emoji prediction in terms of emoji categories and positions. Each tweet in our dataset contains both textual, visual and the user demographic information. We release the dataset and related information upon the publication of this work[2].
- We investigate the correlation between emoji usages and three key aspects. We then propose an mmGRU model to predict the categories and positions of emojis. Our experimental results validate the rationality and effectiveness of the proposed mmGRU model.
- We unveil several interesting findings based on our data, method and experiments, including: 1) users in different geographic regions have different diversity in emoji selection; 2) users in different regions have different understanding on the emotions of the same emojis; 3) emojis in different shapes (i.e. heart shape or face shape) tend to appear at different positions.

The remainder of this paper is organized as follows. First, we introduce the related work in section 2. We then define several notations and formulate the learning tasks in section 3 and discuss data observation in section 4. Next, we describe the proposed emoji prediction method in section 5. Experiments are given in section 6 and section 7 concludes this paper.

[2]http://emoticonprediction.droppages.com/

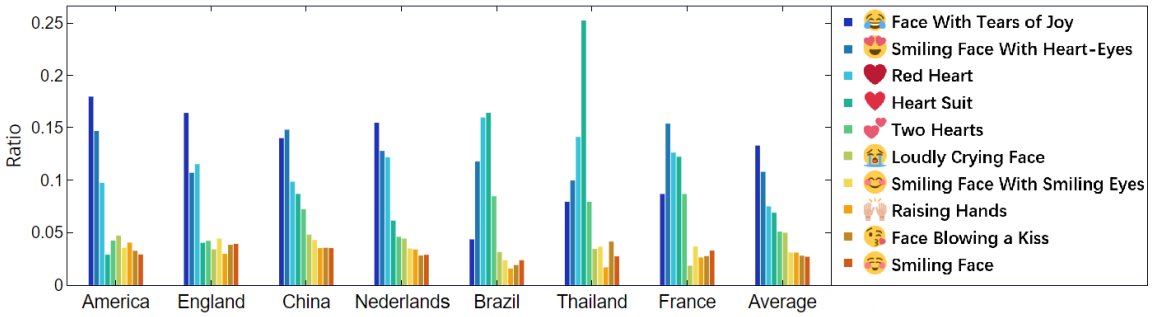
2 RELATED WORK

Social emotion analysis. People often use emojis to express emotions. Research on emotion analysis or prediction has been discussed in many works. [1, 20] describe sentiment analysis methods using textual and visual contents. [21] proposes a new neighborhood classier based on the similarity of two instances described by the fusion of textual and visual features. Recent years, studies have found that user demographics can also help for sentiment analysis for online users. [18] studies the influence of user demographics on image sentiment analysis. [19] combines the textual, visual and the social information of users and propose a user-level emotion prediction method. The above works reveal that adopting multimodality information and user demographics can improve personalized emotion prediction.

Emoji Usages and Recommendation. Traditional works usually study emoji usages through two aspects: treat emojis as sentimental labels, and recommend emojis with textual information. For sentiment analysis, [22] divides emojis into four kinds of emotions and then uses them to study emotion changes of online users. [15] studies the multilingual sentiment analysis using emojis and keywords. [11] labels the sentiment scores of emojis. [6] builds an emoji space model for sentiment analysis. [10] studies the sentiment and semantic misconstrue of emojis on different platforms. [5] studies the intentions and sentiment effects of using emojis during communications. While for emoji recommendation, [8] uses Affective Trajectory Model for emoji category recommendation based on textual information. [16] analyzes the emotion of the textual information and proposes an emoji recommendation method based on the emotive statements of users and their past selections of emojis. The above works construct their recommendation methods mainly using textual information and ignore the complexity of emoji usages such as tweet structure and users' personalized information.

3 PROBLEM FORMULATION

To formulate our problem, we first define some notations.



(a) Regions and Emoji Categories

(b) Emoji Categories and Positions

(c) Regions and Emoji Positions

Figure 2: Observations on the following aspects: (a) Correlation between regions and emoji categories. X-axis represents the seven typical regions and the average. Y-axis represents the ratio of the emojis. The legend represents the top-10 frequently used emojis. (b) Correlations between emoji categories and positions. (c) Correlations between regions and emoji positions.

Definition 1. **Emoji category and position.** The emoji category of a tweet is denoted as $C_i, i \in [0, C)$, where C is the number of emoji categories. As for the emoji position, we define two representations: fine-grained position and coarse-grained position. The fine-grained position of a tweet is denoted as $FP_i, i \in [0, P)$, where P is the max length of the tweets. The coarse-grained position of a tweet is denoted as $CP_i, i \in [0, 3)$, where CP_0, CP_1, CP_2 represent the start, middle and end positions.

Definition 2. **mmGRU network.** The mmGRU network is denoted as $G = (\mathbf{X}, \mathbf{V}, \mathbf{U})$, where $\mathbf{X}$ is $l \times d$ attribute matrix, each row corresponds to a tweet and each column represents one dimension of the textual features. $\mathbf{V}$ is $l \times e$ attribute matrix, each row corresponds to a tweet and each column represents one dimension of the visual features. $\mathbf{U}$ is $l \times f$ attribute matrix, each row corresponds to a tweet and each column represents one dimension of the user demographic features. l represents the number of tweets and d, e, f represent the dimensions of each feature, respectively.

Problem 1. **Learning Task.** The problem we focus on is to predict an emoji for a tweet and a suitable position to put it in. We use a multitask multimodality gated recurrent unit (mmGRU) model G_1 to learn a prediction function f_1 to predict the emoji category and coarse-grained position of every tweet, and a multimodality gated recurrent unit model G_2 to learn a prediction function f_2 to predict the fine-grained position of every tweet, defined as:

$$\begin{aligned} f_1 : G_1 &= (\mathbf{X}, \mathbf{V}, \mathbf{U}) \rightarrow \mathbf{C} \textit{ and } \mathbf{CP} \\ f_2 : G_2 &= (\mathbf{X}, \mathbf{V}, \mathbf{U}, \mathbf{CP}) \rightarrow \mathbf{FP} \end{aligned} \tag{1}$$

Table 1: User number of the select regions

Region	America	England	China	Netherlands
User number	30, 361	4, 763	3, 348	2, 978
Region	Brazil	Thailand	France	
User number	2, 410	1, 423	760	

4 DATA-BASED OBSERVATIONS

The choices of emoji categories and positions depend on a number of complex factors, such as tweet content, tweet structure and user demographics. Since people use emojis as the direct reflection on tweet contents, we mainly focus on obtaining a series of observations to further reveal whether tweet structure and user demographics have influences on emoji usage.

To obtain the observations, we first construct a benchmark emoji dataset from Twitter which contains 164,880 tweets of the 35 most-frequently used (0.62%-15.7%) emojis for experimental analysis. Each tweet contains textual, visual, user demographic information and only one emoji. For each of the selected emojis, there are more than 1,000 tweets.

4.1 Observations on Emoji Categories and Positions

Figure 2(b) shows the correlation between the top 10 frequently used emojis and their probabilities of appearing at different positions. The average position ranges from 0 (the beginning of the tweets) to 1 (the end of the tweets). The three rows represent the percentage of emojis that appeared at the start, middle and end of the tweets.

On the average, 70% of the emojis are used at the end of the tweets and only 2.6% are used at the beginning of the tweets. “Heart Suit" and “Red Heart" are more likely to be used in the front part of the tweets. Comparing with other emojis, “Raising Hands" and “Face with Tears of Joy" are more likely to be used at the end of tweets.

4.2 Observations on Users’ Demographics and Their Choices of Emojis

Cultural background will affect the habit of using emojis. Geographic region information has a great influence on the cultural background of users. Therefore, we take region information as an example to represent the demographics.

Category. We select seven regions that with the largest number of users which are shown in Table 1. The selected regions are America, China, England, France, Netherlands, Brazil and Thailand. Figure 2(a) shows the correlation between the regions and the top 10 frequently used emojis. We find that people in different regions have their preferred emojis. For example, “Heart Suit", which is used in card games for the heart suit, is more likely to be used by users in France, Brazil, and Thailand but rarely by the users in America. “Red Heart", which looks like “Heart Suit", is used for expressions of love according to the explanation in Emojipedia[3]. This emoji is more popular in Brazil and less popular in China. Users in America, China and Nederlands are more likely to use “Loudly Crying Face", while users in France almost never use that emoji, which means that users in France rarely share their pessimistic emotion online. The region closest to the average (emoji distributions for all users) is Netherlands.

Position. Figure 2(c) shows the correlations between the regions and positions of emojis. We count the percentage of emojis that appeared at the start, middle and end of the tweets and the average emoji position for every region. Users in Brazil and Thailand are more likely to employ emojis in the front part of the tweets. Users in America and China are more likely to employ emojis at the end of tweets, and Chinese users barely employ emojis at the start of tweets.

Summarization. The observations can be summarized as follows:

- There exists correlation between emoji categories and positions. In particular, most of the emojis are mainly used at the end of the tweets and some are more likely to be used in the front and the middle part of the tweets, such as “Heart Suit" and “Red Heart".
- Regional culture can affect the users’ habits of using emojis, which suggests that demographics such as region information should be considered to improve emoji prediction.

5 PROPOSED METHOD

Inspired by the observations, we propose a multitask multimodality gated recurrent unit (mmGRU) model to quantitatively describe the correlations between emojis and tweet content, tweet structure, user demographics. We use GRU[3] as the framework because it can overcome the problem of gradient vanishing compared with recurrent neural network (RNN). The multimodality function is used to incorporate multimodality information such as text, image and user demographics, while the multitask framework helps leverage the strong correlation between emoji categories and their positions to improve the performance of both tasks.

[3]https://emojipedia.org/

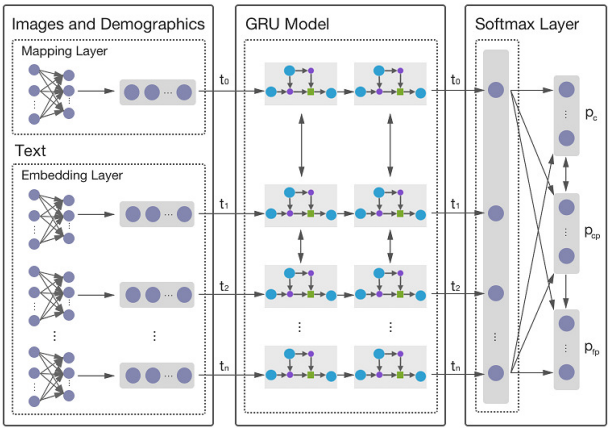


Figure 3: One slice of the proposed mmGRU model for position prediction

An overview of the proposed method for emoji prediction is shown in Figure 1. In the first step, we extract the textual, visual and the user demographic features from a tweet. Then, the three kinds of features are sent into the proposed mmGRU model to learn an output representation. Next, we use a softmax classifier to obtain the predicted results of emoji categories and coarse-grained positions together according to the their correlation as mentioned in section 4. As for fine-grained position, we use another softmax classifier and the results of coarse-grained position to get the predicted results. Finally, we can obtain the target emojis and their positions in the raw tweets based on the well-trained model.

5.1 Model structure

We explain our model in Figure 3. The rectangular box represents a cell of GRU. Compared with LSTM cell, it only contains two gates: a reset gate and an update gate. The reset gate determines how to combine the new input with the previous memory, and the update gate defines how much of the previous memory to keep around. The equations of the GRU cell are as follows:

$$\begin{aligned} r_t &= \sigma\left(W_{xr}x_t + W_{hr}h_{t-1}\right) \\ z_t &= \sigma\left(W_{xz}x_t + W_{hz}h_{t-1}\right) \\ \widetilde{h_t} &= tanh\left(W_{xr}x_t + W_{ho}h_{t-1}\right) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \end{aligned} \tag{2}$$

where r_t is the reset vector, z_t is the update vector, $\widetilde{h_t}$ is the candidate output vector, h_t is the output vector, $W_{xr}, W_{hr}, W_{xz}, W_{hz}$ are the weight matrix parameters, σ is the sigmoid function and tanh is a hyperbolic tangent.

For textual features $\mathbf{X} = x_i, i \in [1, t]$, we use word embedding method[4] to convert the one-hot features into low-dimensional vectors. Then we get the input vectors $\widetilde{\mathbf{X}} = \widetilde{x_i}, i \in [1, t]$.

[4]embedding_lookup function in Tensorflow

For visual features and user demographic features, we use the joint learning method proposed in [17]. We use W_v and b_v to convert the visual features $\mathbf{V}$ and user demographic features $\mathbf{U}$ into textual feature space. After the new vector is learned, we treat it as the initial state of the model before we input each word in every tweet, which means we send the image into the model at time 0. We combine the vector $\widetilde{x_0}$ with the textual features $\widetilde{x_i}, i \in [1, t]$ we get from the embedding layer.

$$\widetilde{x_0} = W_v\,(\mathbf{V} + \mathbf{U}) + b_v \tag{3}$$

where W_v is the weight matrices and b_v is the bias vector.

The hidden layer is composed of a bidirectional GRU (BI-GRU) hidden layer, which is used for building output vectors $y_i, i \in [0, t]$ of input features.

$$y_0, y_1, \cdots, y_t = BI - GRU\,(\widetilde{x_0}, \widetilde{x_1}, \cdots, \widetilde{x_t}) \tag{4}$$

We use max pooling operation to convert the output vectors $y_i, i \in [0, t]$ into the final output vector $\mathbf{o}$.

$$\mathbf{o} = MaxPooling\,(y_0, y_1, \cdots, y_t) \tag{5}$$

We use a linear translation and softmax regression to gain the probability distribution of categories and coarse-grained positions. Then we use the results of coarse-grained position to calculate the probability distribution of fine-grained positions.

$$\begin{aligned} \mathbf{o}_c^s &= W_c \times \mathbf{o} + b_c \\ p_{ci} &= \frac{exp(\mathbf{o}_{c\,i}^s)}{\sum_{n=1}^{C} exp(\mathbf{o}_{c\,n}^s)}, i \in [0, C) \end{aligned} \tag{6}$$

$$\begin{aligned} \mathbf{o}_{cp}^s &= W_{cp} \times (\mathbf{o} + p_c) + b_{cp} \\ p_{cpi} &= \frac{exp(\mathbf{o}_{cp\,i}^s)}{\sum_{n=1}^{C} exp(\mathbf{o}_{cp\,n}^s)}, i \in [0, 3) \end{aligned} \tag{7}$$

$$\begin{aligned} \widetilde{\mathbf{o}_c^s} &= \widetilde{W_c} \times (\mathbf{o} + p_{cp}) + \widetilde{b_c} \\ \widetilde{p_{ci}} &= \frac{exp(\widetilde{\mathbf{o}_{c\,i}^s})}{\sum_{n=1}^{C} exp(\widetilde{\mathbf{o}_{c\,n}^s})}, i \in [0, C) \end{aligned} \tag{8}$$

$$\begin{aligned} \mathbf{o}_{fp}^s &= W_{fp} \times (\mathbf{o} + p_{cp}) + b_{fp} \\ p_{fpi} &= \frac{exp(\mathbf{o}_{fp\,i}^s)}{\sum_{n=1}^{C} exp(\mathbf{o}_{fp\,n}^s)}, i \in [0, P) \end{aligned} \tag{9}$$

where $\widetilde{p_{ci}}/p_{cpi}/p_{fpi}$ is the probability for the i-th category/ coarse-grained position/ fine-grained position. p_{ci} is intermediate results for category. C is the number of categories. $\mathbf{o}_c^s, \mathbf{o}_{cp}^s, \widetilde{\mathbf{o}_c^s}, \mathbf{o}_{fp}^s$ are the output vectors after linear translation. W_c, b_c, W_{cp}, b_{cp}, W_{fp}, b_{fp}, $\widetilde{W_c}$ and $\widetilde{b_c}$ are the parameters of softmax regression that need to be learned.

5.2 Model learning

. As for category and coarse-grained position, we calculate the loss function by incorporating the loss values of both two prediction tasks and train to minimize the value of the loss function J_1. As for the fine-grained position, we calculate the loss function J_2 alone. The training process uses Adam optimization [7] to perform the classifier-to-neural network feedback and feeds back into the input word vector layer.

Algorithm 1 mmGRU Model for Category and Coarse-grained Position

Input: $I = \{\mathbf{X}, \mathbf{V}, \mathbf{U}\}$, a preprocessed feature matrix.
Output: Final parameter $\theta_1 = \{W_{cp}, b_{cp}, \widetilde{W_c}, \widetilde{b_c}\}$, the parameter after the training process.

Initialize model parameters θ_1, α_1
repeat
  $W_{cp} = W_{cp} - \alpha_1 \frac{\delta}{\delta W_{cp}} J_1(W_{cp}, b_{cp})$
  $b_{cp} = b_{cp} - \alpha_1 \frac{\delta}{\delta b_{cp}} J_1(W_{cp}, b_{cp})$
  $\widetilde{W_c} = \widetilde{W_c} - \alpha_1 \frac{\delta}{\delta \widetilde{W_c}} J_1(\widetilde{W_c}, \widetilde{b_c})$
  $\widetilde{b_c} = \widetilde{b_c} - \alpha_1 \frac{\delta}{\delta \widetilde{b_c}} J_1(\widetilde{W_c}, \widetilde{b_c})$
until convergence
return θ_1

Algorithm 2 mmGRU Model for Fine-grained Position

Input: $I = \{\mathbf{X}, \mathbf{V}, \mathbf{U}, \mathbf{CP}\}$, a preprocessed feature matrix.
Output: Final parameter $\theta_2 = \{W_{fp}, b_{fp}\}$, the parameter after the training process.

Initialize model parameters θ_2, α_2
repeat
  $W_{fp} = W_{fp} - \alpha_2 \frac{\delta}{\delta W_{fp}} J_2(W_{fp}, b_{fp})$
  $b_{fp} = b_{fp} - \alpha_2 \frac{\delta}{\delta b_{fp}} J_2(W_{fp}, b_{fp})$
until convergence
return θ_2

$$\begin{aligned} \theta_1{}^* &= \arg\min_{\theta_1} J_1(\widetilde{W_c}, \widetilde{b_c}, W_{cp}, b_{cp}) \\ \theta_2{}^* &= \arg\min_{\theta_2} J_2(W_{fp}, b_{fp}) \end{aligned} \tag{10}$$

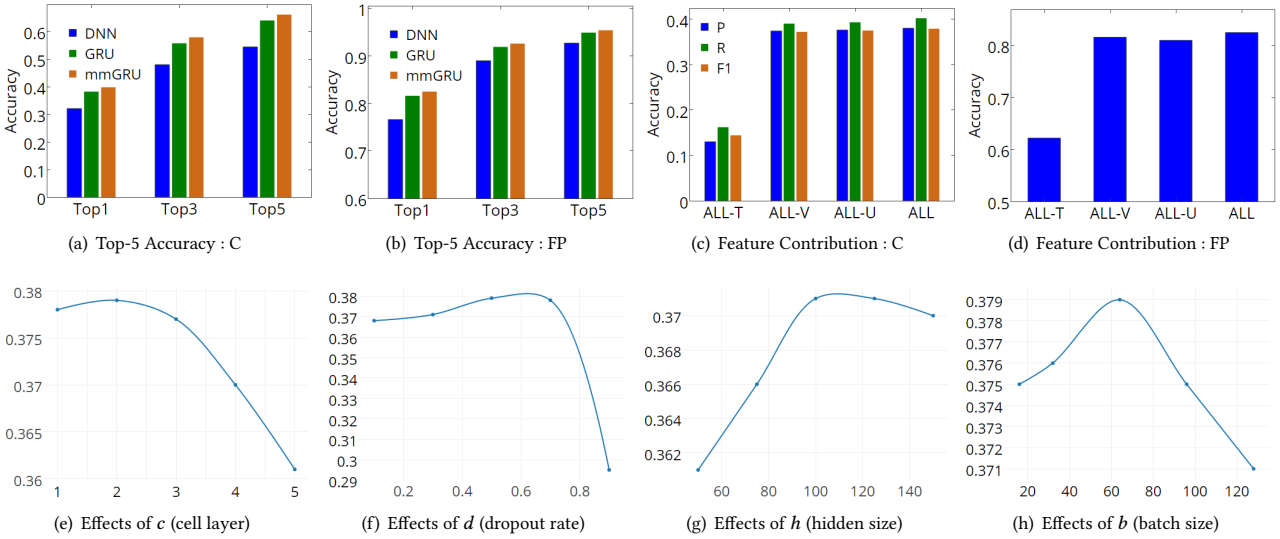
6 EXPERIMENTS

6.1 Experimental Setup

Data set. We use the dataset collected in section 4. It contains 164,880 Twitter tweets. And each tweet contains textual, visual, regional information and an emoji belongs to the 35 most frequently used emojis. The dataset is collected from the raw dataset that mentioned in [13], which is streamed from Twitter API using a set of keywords related to YouTube and its videos. For the experimental setup, we design a prediction task for the select 35 emojis and their positions in tweets.

Comparison methods. We conduct performance comparison experiments to demonstrate the effectiveness of our model. We select five methods as follows: Random selection, Logistic Regression (LR), Support Vector Machine (SVM), Deep Neural Network (DNN), Gated Recurrent Unit (GRU) and the proposed model.

- **Random**. We adopt the random selection method as the first baseline method. The results of this method are equal to the max proportion of all the prediction choices.



(a) Top-5 Accuracy : C (b) Top-5 Accuracy : FP (c) Feature Contribution : C (d) Feature Contribution : FP

(e) Effects of c (cell layer) (f) Effects of d (dropout rate) (g) Effects of h (hidden size) (h) Effects of b (batch size)

Figure 4: Experimental results: (a) top-5 accuracy of category; (b) top-5 accuracy of fine-grain position; (c) feature contribution analysis of category; (d) feature contribution analysis of fine-grain position; (e) effects of the cell layer number; (f) effects of the dropout rate; (g) effects of the number of hidden states; (h) effects of the batch size.

- **LR**. Logistic regression is commonly used in classification problems. Here we use the scikit-learn [12] implementation to build our baseline method.
- **SVM**. SVM is a frequently-used method of solving classification problems. Here we use LIBSVM[5] to build it.
- **DNN**. It has been proved that DNN can achieve good performance in multimodality classification [9]. We use the DNN model it mentioned as the baseline method in our experiments.
- **GRU**. Basic GRU model[2] that trains emoji categories and positions separately.
- **The proposed method**. The mmGRU method we propose in this paper.

Metrics. To quantitatively evaluate the category and position prediction performance, we use micro *Precision*, *Recall* and *F1-value*[6] as metrics, which are calculated by the weighted average method. Because our goal is a multi-classification problem. We also provide *Top-k accuracy* as another metric for category prediction performance.

We implement, train and evaluate our method on Tensorflow[7]. We perform five-fold cross validation to obtain the average prediction performance.

6.2 Feature Extraction

Textual Features. First, we load the words of the Twitter corpus and translate each word into a one-hot vector. Then we replace the words that appeared less than 5 times with a symbol called "<rare>". Next we formulate all tweets into the same length by adding a symbol called "<blank>" to the end of tweets.

Visual Features. We use VGG-19 [14] to help build our visual feature representations. VGG-19 is a deep convolutional network with 19 layers which is designed for feature extraction tasks and classification tasks of images. We choose the second to the last layer as the feature representation since it can represent the global information of the images. Then we obtain a vector with 4,096 dimensions from each picture as the visual features.

User Demographic Features. Considering the users' privacy, we only choose region information as an example of the user demographic features in our experiments. We collect the region features such as the country information of the users from every tweet and then transform them into one-hot vectors.

6.3 Results and Analysis

Performance analysis. First, we aim to validate if our model is effective for the prediction tasks. We compare the proposed model with several baseline methods.

Table 2 lists the average prediction results of the mentioned models. The proposed mmGRU model clearly shows the best performance than other models. In terms of *F1-value* for category, our model achieves 14.4% improvement compared with LR, 13.3% improvement compared with SVM, 7.5% improvement compared with DNN and 0.8% improvement compared with GRU. As for coarse-grained (fine-grained) position, our model achieves 6.6% (7.9%) improvement compared with LR, 5.0% (6.7%) improvement compared with SVM, 3.2% (5.9%) improvement compared with DNN and 0.8% (0.9%) improvement compared with GRU on average. We use T-test[8] to confirm if there is a significant difference between our method and the baseline methods. The results show that the P-value[9] is less than 0.05, which means our method has a significant improvement over baseline methods. Specifically, we list the F1-value of the top 10 mostly used emojis in Table 3. The results show that our model can achieve the best performance of most emojis. All of the above experiments confirm that the multimodal

[5] A library for SVM designed by Chang and Lin
[6] https://en.wikipedia.org/wiki/F1_score
[7] www.tensorflow.org
[8] https://en.wikipedia.org/wiki/Statistical_significance
[9] https://en.wikipedia.org/wiki/P-value

Table 2: Comparison of results using different models

	Category			C-Position	F-Position
Metrics	Precision	Recall	F1-score	F1-score	F1-score
Random	0.157	0.157	0.157	0.700	0.700
LR	0.321	0.316	0.235	0.785	0.742
SVM	0.325	0.326	0.246	0.801	0.754
DNN	0.354	0.379	0.304	0.819	0.762
GRU	0.375	0.390	0.371	0.843	0.812
mmGRU	**0.380**	**0.402**	**0.379**	**0.851**	**0.821**

Table 3: Category performance on the top 10 frequently used emojis (in terms of F1-value)

Emoji	Random	LR	SVM	DNN	GRU	mmGRU
😂	0.157	0.699	0.697	0.702	0.724	**0.730**
😍	0.141	0.412	0.427	0.448	0.524	**0.530**
❤️	0.114	0.268	0.296	0.348	**0.479**	0.465
❤️	0.060	0.349	0.361	0.415	0.457	**0.487**
💕	0.049	0.210	0.255	0.265	**0.360**	0.328
😭	0.040	0.078	0.086	0.189	**0.285**	0.282
😊	0.038	0.030	0.034	0.077	0.195	**0.202**
🙌	0.034	0.074	0.075	0.243	0.360	**0.368**
😘	0.033	0.022	0.056	0.238	0.308	**0.342**
☺️	0.031	0.022	0.039	0.148	0.245	**0.255**

multitask framework and the GRU network are effective in modeling emoji usage. We also notice that for "Red Heart", "Two Hearts" and "Loudly Crying Face", the performance of our model are worse than the performance of GRU model. It is probably because the correlation between the categories and positions of these emojis are not significant.

Meanwhile, considering emoji recommendation, top-k most likely emojis are very useful results. We also design an experiment of top-k accuracy metrics. In our task, the top-k emojis are selected by the probability distribution of emojis in the prediction results. We set k to 1, 3, 5, and compare our method with DNN and GRU, which achieves better performance than the other three baseline methods. As Figure 4(a) shows, our method achieves 11.7% improvement compared with DNN, 2.1% improvement compared with GRU and reach an accuracy of 66.1% for top-5 emoji category recommendation. As Figure 4(b) shows, our method achieves 2.7% improvement compared with DNN, 0.6% improvement compared with GRU and reach an accuracy of 95.4% for top-5 emoji position recommendation.

Feature contribution analysis. In our work, we utilize the textual, visual and user demographic features and send them into an mmGRU model for emoji prediction tasks. To investigate whether these features benefit the prediction tasks, we investigate the contributions of every kind of feature. Every time, we take each of the features out from the primitive model while keeping the other features and then examine the performance.

Figure 4(c) and 4(d) shows the category and fine-grained position performance of different feature combinations. The model involving all factors achieves the best performance both in category and fine-grained position. Textual features achieve the most contribution (+23.5% in *F1-value* for category and +20.3% in *F1-value* for position) among all features. For the last two kind of features, visual feature (+0.7% in *F1-value*) achieves greater contribution than user demographic feature (+0.4% in *F1-value*) for category prediction. On the contrary, user demographic feature (+1.5% in *F1-value*) achieves greater contribution than visual feature (+0.9% in *F1-value*) for position prediction. The above results verify the effectiveness of the multimodal features and the multimodal function of our mmGRU model.

Parameter sensitivity analysis. We conduct experiments for parameter adjustment in our training process. We show how the changes of parameters in mmGRU affect the performance of emoji category prediction by comparing the average performance in five-fold cross validation.

- Cell layer, the number of cell layers. Visualized in Figure 4(e), as the number of layers in mmGRU increases, the performance turns better at first and then declines. The performance achieves the highest value when the number of layers is 2.
- Dropout rate, the dropout probability of the input. Figure 4(f) shows the performance of the dropout rate. As the dropout rate increases, the performance turns also better at first and then declines. The performance achieves the highest value when the dropout rate is 0.5.
- Hidden size, the number of the hidden states. Figure 4(g) shows the performance of it. The performance achieves the highest value when the hidden size is 100.
- Batch size. The performance shown in Figure 4(h) illustrates that when batch size is 64, the performance achieves the highest value.

Based on the parameter sensitivity analysis, we select the final parameters as follows: In the cell of GRU, we adopt two cell layers. The batch size is 64, dropout rate is 0.5 and hidden size is 100. Our experiments are conducted on a X64 machine with K80 GPU and 128G RAM.

6.4 Case Study

In this part, we would like to give three interesting case studies of our data, method and experiments in Figure 5.

People in different regions have different diversity in emoji selection. We find that the accuracy of the predictions is different in different regions, indicating that people in different regions have different diversity in emoji selection. Users in Brazil are the most consistent in using emojis while users in China are the most diverse in using emojis.

Users in different regions have different understandings on the emotions of the same emojis. "Face with Tears of Joy" has both positive or negative meanings. We analyze the expressions of positive, negative and neutral emotions of this emoji in different regions by calculating the sentiments (the composition of positive, neutral and negative emotion and the sentiment score)[10] of textual information in tweets. It shows that Brazilians prefer to express their positive emotions with it while French prefer to express their negative emotions with it.

Emojis in different shapes (i.e. heart shape or face shape) tend to appear at different positions. Heart emojis are more likely to appear in the middle of the sentences, while face emojis are more likely to appear at the end of the sentences. It shows

[10] http://www.nltk.org/api/nltk.sentiment.html

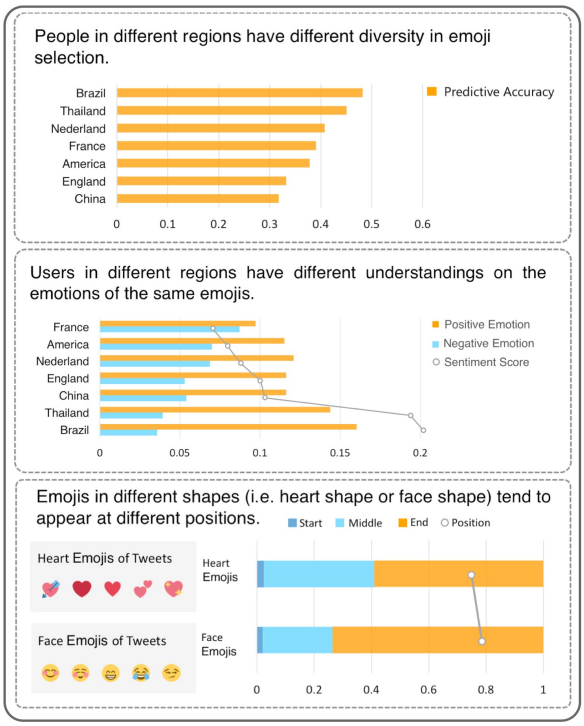


Figure 5: Three interesting case studies of our data, method and experiments. Top: prediction accuracy of different regions. Middle: sentiment understanding of "Face with Tears of Joy" in different regions. Bottom: distributions of positions on face emojis and heart emojis.

that heart emojis are often used to express the emotions locally in the sentence, while the face emojis are often used to enhance the emotions of the sentence globally.

7 CONCLUSION

In this paper, we examine the correlation between emoji usages and tweet content, tweet structure, user demographics. We then propose an mmGRU model for predicting emoji categories and positions motivated by the observations. The multimodality function is used to incorporate multimodality information such as text, image, and user demographics, while the multitask framework helps leveraging the strong correlation between emoji categories and their positions for improving the performance of both tasks. Extensive experiments have shown the effectiveness of the proposed method. The goal of our work is to generate reasonable emojis for the input of online users which can make the delivery of semantics more accurate and make computer-mediated communication more effective. Our work has many concrete applications, including improving user-to-user as well as user-to-chatbot interactions in social networks by building emoji recommendation systems or designing emoji input methods for online users. We plan to release the benchmark emoji-labeled dataset to facilitate more research in this area.

8 ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Plan (2016YFB1001200), the Innovation Method Fund of China (2016IM010200), the National Natural and Science Foundation of China (61521002), and the (US) National Science Foundation (1704309). We would also like to thank Tiangong Institute for Intelligent Computing, Tsinghua University for its support.

REFERENCES

[1] Donglin Cao, Rongrong Ji, Dazhen Lin, and Shaozi Li. 2014. A cross-media public sentiment analysis system for microblog. *Multimedia Systems* (2014), 1–8.
[2] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Computer Science* (2014).
[3] Bhuwan Dhingra, Hanxiao Liu, William W. Cohen, and Ruslan Salakhutdinov. 2016. Gated-Attention Readers for Text Comprehension. (2016).
[4] Thomas Dimson. 2015. Emojineering Part 1: Machine Learning for Emoji Trends. http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji. (2015).
[5] Tianran Hu, Han Guo, Hao Sun, Thuyvy Thi Nguyen, and Jiebo Luo. 2017. Spice up Your Chat: The Intentions and Sentiment Effects of Using Emoji. In *International AAAI Conference on Web and Social Media*. 102–111.
[6] Fei Jiang, Yiqun Liu, Huanbo Luan, Min Zhang, and Shaoping Ma. 2014. *Microblog Sentiment Analysis with Emoticon Space Model.* Springer Berlin Heidelberg. 76–87 pages.
[7] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *Computer Science* (2014).
[8] Wei-Bin Liang, Hsien-Chang Wang, Yi-An Chu, and Chung-Hsien Wu. 2014. Emoticon recommendation in microblog using affective trajectory model. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 1–5.
[9] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *ACM International Conference on Multimedia*. 507–516.
[10] Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "Blissfully happy" or "ready to fight": Varying Interpretations of Emoji. In *International AAAI Conference on Web and Social Media*.
[11] Petra Kralj Novak, Jasmina SmailoviÄĞ, Borut Sluban, and Igor MozetiÄĮ. 2015. Sentiment of Emojis. *Plos One* 10, 12 (2015).
[12] Fabian Pedregosa, Ga Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2013. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 10 (2013), 2825–2830.
[13] Marian Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. 2017. Expecting to be HIP: Hawkes Intensity Processes for Social Media Popularity. In *International Conference on World Wide Web*. 735–744.
[14] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science* (2014).
[15] Georgios S. Solakidis, Konstantinos N. Vavliakis, and Pericles A. Mitkas. 2014. Multilingual Sentiment Analysis Using Emoticons and Keywords. In *Ieee/wic/acm International Joint Conferences on Web Intelligence*. 102–109.
[16] Yuki Urabe, Rafal Rzepka, and Kenji Araki. 2014. Emoticon Recommendation System to Richen Your Online Communication. *International Journal of Multimedia Data Engineering & Management* 5, 1 (2014), 20.
[17] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.
[18] Boya Wu, Jia Jia, Yang Yang, Peijun Zhao, and Jie Tang. 2015. Understanding the emotions behind social images: Inferring with user demographics. In *IEEE International Conference on Multimedia and Expo*. 1–6.
[19] Yun Yang, Peng Cui, Wenwu Zhu, and Shiqiang Yang. 2013. User interest and social influence based emotion prediction for individuals. In *ACM International Conference on Multimedia*. 785–788.
[20] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia. In *ACM International Conference on Web Search and Data Mining*.
[21] Yaowen Zhang, Lin Shang, and Xiuyi Jia. 2015. Sentiment Analysis on Microblogging by Integrating Text and Image Features. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 52–63.
[22] Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. 2012. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1528–1531.