THE UNIVERSITY OF RHODE ISLAND

Emoji Prediction Using Bidirectional LSTM

- Arup Mazumder
- Piriyankan Kirupaharan

THINK BIG WE DO

Introduction

- Emoticon is a symbol that represents a face expression, like 😊 . It enables the users to convey feelings, moods, and emotions and enhances written communication with non-verbal cues.

- Text suggestion is a feature that improves the user experience. Similarly we can have emoji suggestion for given text that will improve the user experience.

Goal

- We wanted to build a machine learning model that will accurately predict contextual emoji on a given text.

Try our demo: http://18.141.220.45/

Data

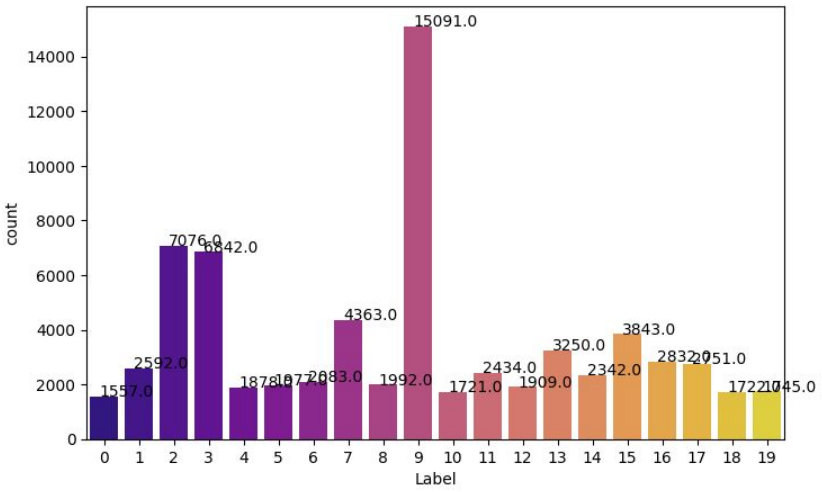
- Twitter Emoji data consists of 70,000 records with 20 classes.
- Data consist of special characters and tagging usernames. So we cleaned the data to remove those specific words from the sentence.

Before preprocessing

	TEXT	Label
0	Vacation wasted ! #vacation2017 #photobomb #ti...	0
1	Oh Wynwood, you're so funny! : @user #Wynwood ...	1
2	Been friends since 7th grade. Look at us now w...	2
3	This is what it looks like when someone loves ...	3
4	RT @user this white family was invited to a Bl...	3

After preprocessing

	Text	Label
0	vacation wasted vacation photobomb tired vaca...	0
1	oh wynwood you're so funny wynwood art itwas...	1
2	been friends since th grade look at us now we ...	2
3	this is what it looks like when someone loves ...	3
4	rt this white family was invited to a black b...	3



THINK BIG WE DO℠

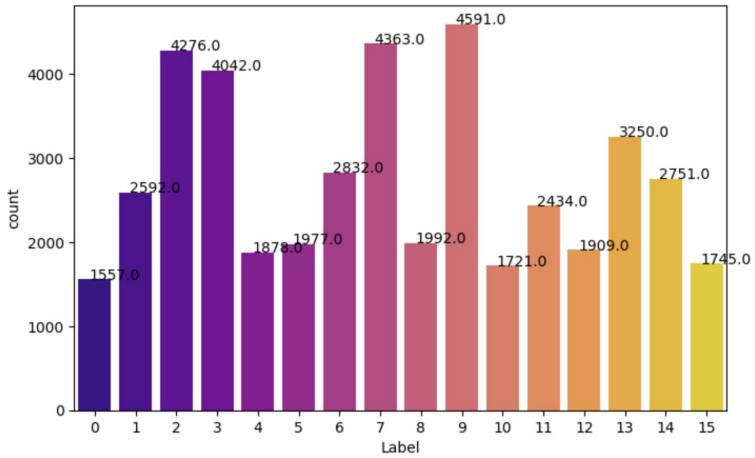
THE UNIVERSITY OF RHODE ISLAND

Try our demo: http://18.141.220.45/

Data

```
18,💜,18 (love, support, close bonds, and admiration for things that have some relation
15,💕,15 love
14,💙,14 love, support, admiration, happiness, and excitement
6,📷,6 classic camera
```

- Since the data is not distributed normally, we removed few records randomly (Eg: records from class 9).
- Some of the emojis had the same meaning, so we removed similar emoji classes. (Eg: 💜,💕,💙,📷)
- After doing all the preprocessing, we got 43, 910 data points with 16 classes.
- We split the data such that training data is 80% and testing data is 20%

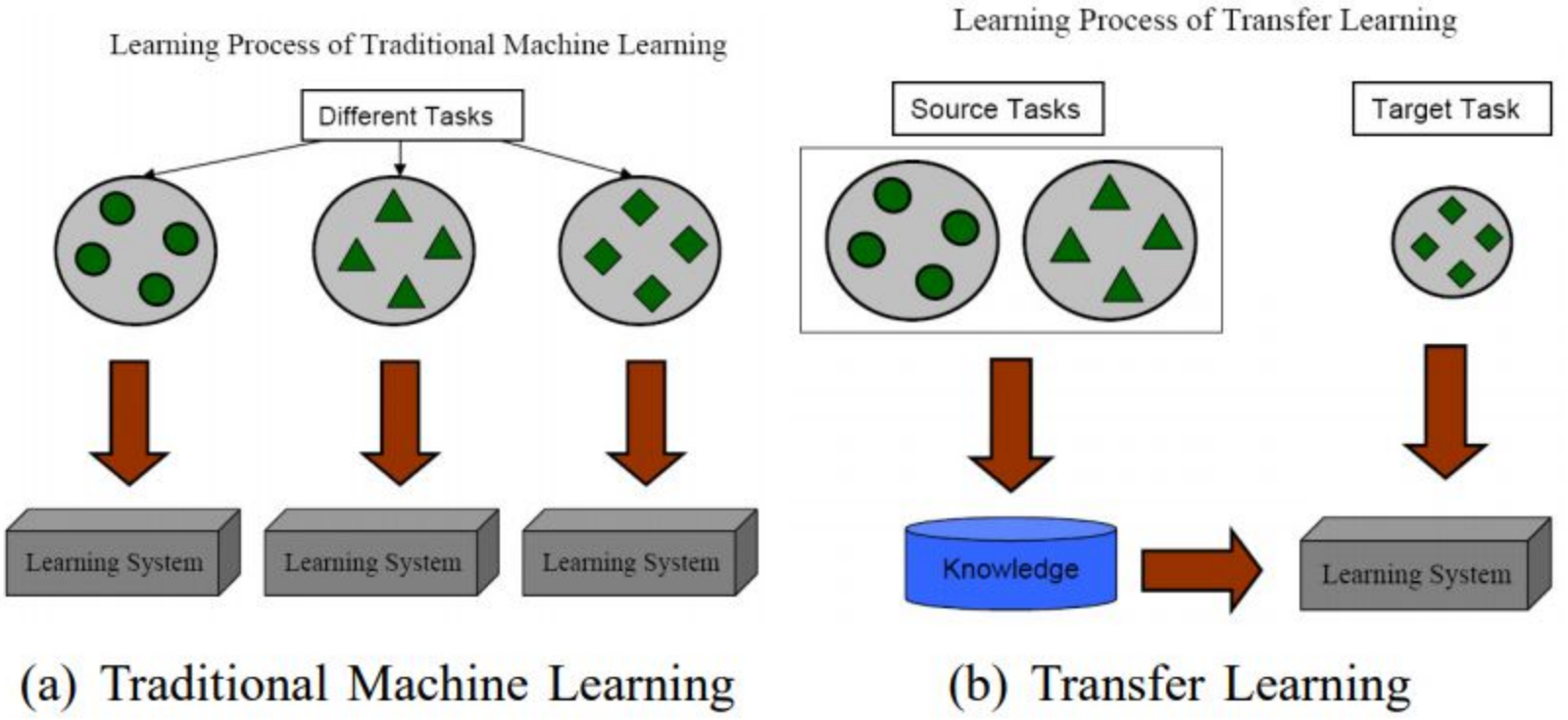


Data

Class	Emoji	Emoji Name/Meaning
0	😜	Fun, joking, or cheeky
1	📸	Photography
2	😍	Love or adoration
3	😂	Extreme happiness or laughter
4	😉	Joking or being cheeky
5	🎄	Christmas
6	😎	Cool or confident
7	🔥	Hot or excellent

Class	Emoji	Emoji Name/Meaning
8	😘	Kissing someone, or general expression of love
9	❤	Gratitude, love, happiness, hope
10	😁	Glowing, beaming happiness
11	🇺🇸	Country flag (USA)
12	✴	Snowflake
13	✨	Positive, happy, or celebration
14	😊	Positive or happy
15	💯	100 percent approval

Experiment: Transfer Learning



Source: https://lisaong.github.io/mldds-courseware/03_TextImage/transfer-learning.slides.html

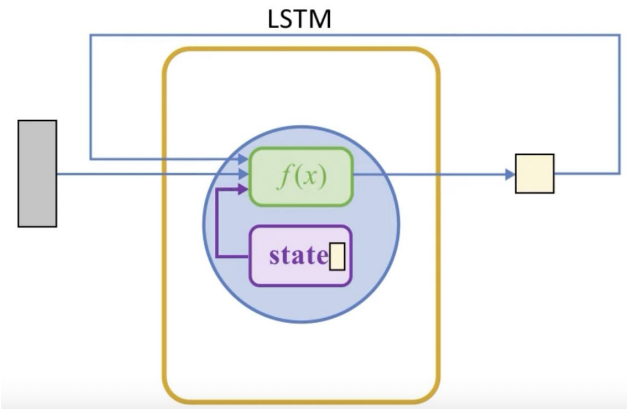
Experiment: Glove6B50D

- Words to vectors: The model is available as text file **Glove 6B 50D**.
- Each word will have 50 values.
- We restricted the maximum number of words in a sentence can be of 20 words.
- We used Glove vectors to convert our words to embedded vectors. So our input matrix will be of size (43,910*20*50)

```
print(embeddings_index['happy'])
print(embeddings_index['happy'].shape)
✓ 0.2s
[ 0.092086  0.2571   -0.58693  -0.37029   1.0828   -0.55466  -0.78142
  0.58696  -0.58714   0.46318  -0.11267   0.2606   -0.26928  -0.072466
  1.247     0.30571   0.56731   0.30509  -0.050312 -0.64443  -0.54513
  0.86429   0.20914   0.56334   1.1228   -1.0516   -0.78105   0.29656
  0.7261   -0.61392   2.4225    1.0142   -0.17753   0.4147   -0.12966
 -0.47064   0.3807    0.16309  -0.323    -0.77899  -0.42473  -0.30826
 -0.42242   0.055069  0.38267   0.037415 -0.4302   -0.39442   0.10511
  0.87286 ]
(50,)
```

Experiment

- Long-Short Term Memory(LSTM): Is a type of Recurrent Neural Network(RNN), designed for application where input is an ordered sequence where information from earlier in the sequence is used.



- Used LSTM:
  - Single Layer LSTM
  - Two Layers LSTM
  - Bidirectional LSTM

Experiment: Single Layer LSTM

- 1 LSTM layer with 512 units
- Dropout layer
- Dense Layer with 128 units
- Output layer: Softmax function with 16 units

Try our demo: http://18.141.220.45/

Experiment: Two Layers LSTM

- 1 LSTM layer with 512 units
- Dropout layer
- 1 LSTM layer with 256 units
- Dropout layer
- Dense Layer with 128 units
- Output layer: Softmax function with 16 units

Try our demo: http://18.141.220.45/

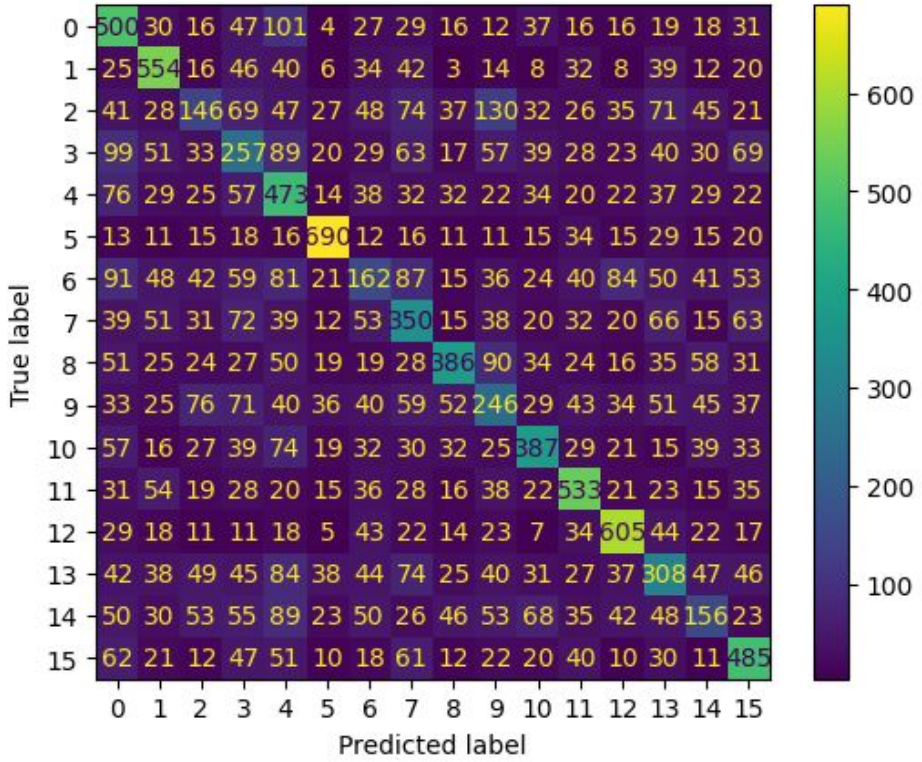
Experiment: Bidirectional LSTM

- 1 bidirectional LSTM layer with 1024 units
- Dropout layer
- 1 bidirectional LSTM layer with 512 units
- Dropout layer
- Dense Layer with 128 units
- Output layer: Softmax function with 16 units

Try our demo: http://18.141.220.45/

Results (Single Layer LSTM)

- Precision : 0.41
- Recall : 0.42
- F1 score : 0.42
- Accuracy: 0.42



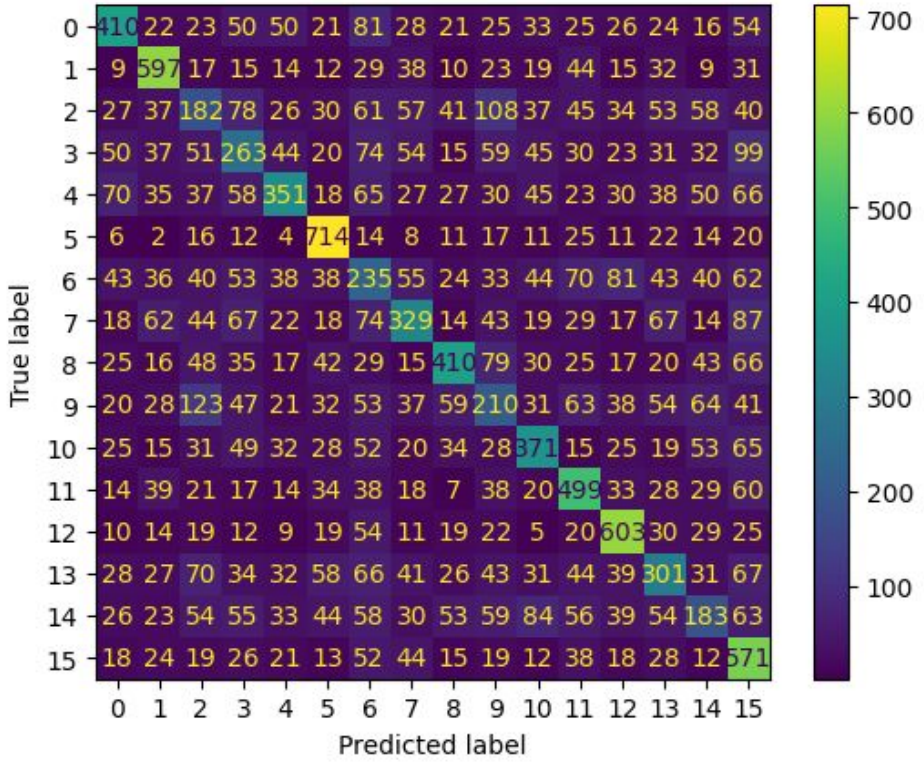
THINK BIG WE DO℠

THE UNIVERSITY OF RHODE ISLAND

Try our demo: http://18.141.220.45/

Results (Two Layers LSTM)

- Precision : 0.41
- Recall : 0.42
- F1 score : 0.41
- Accuracy: 0.42



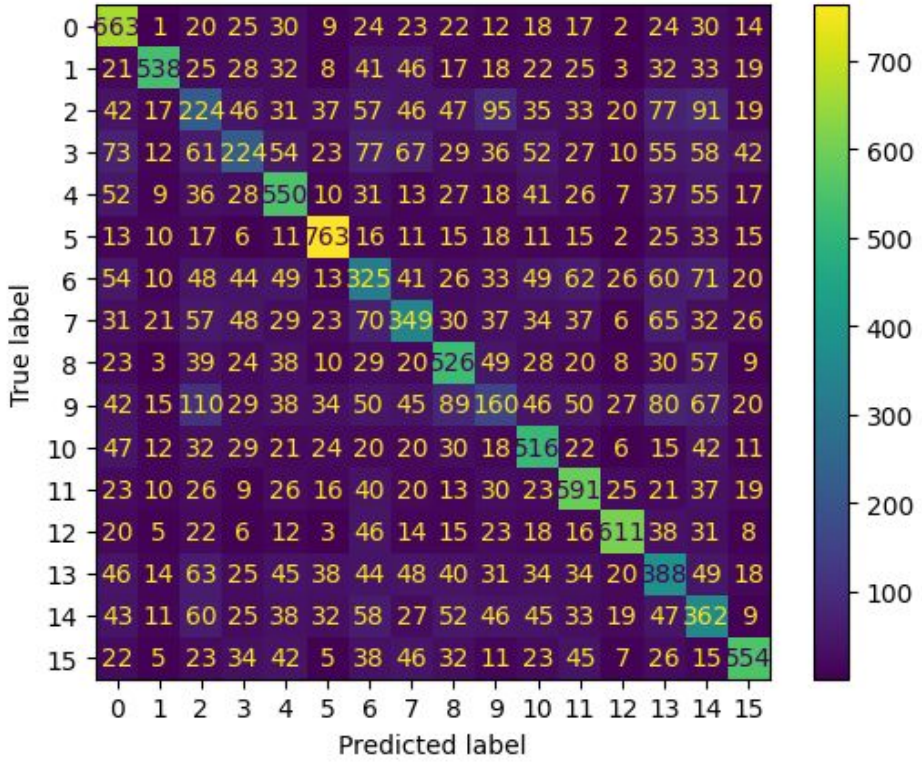
THINK BIG WE DO℠

THE UNIVERSITY OF RHODE ISLAND

Try our demo: http://18.141.220.45/

Results (Bidirectional LSTM)

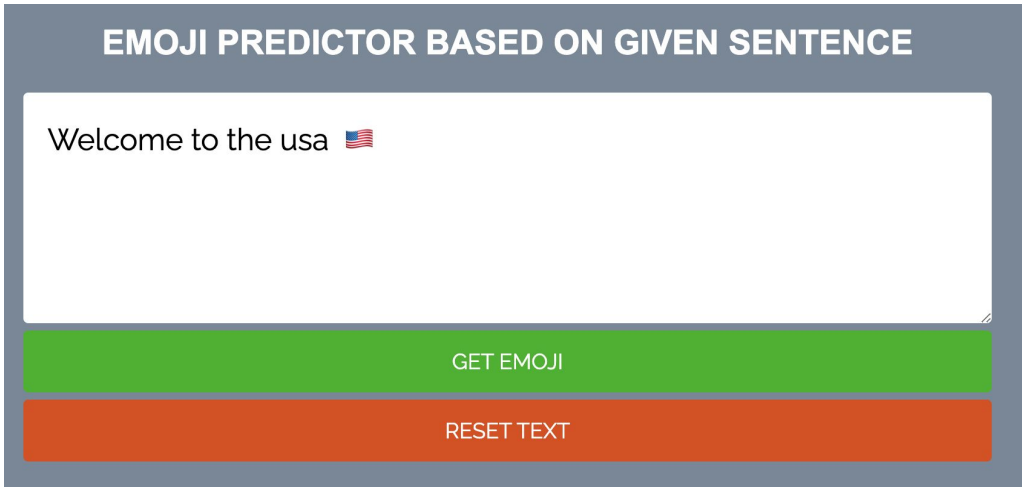
- Precision : 0.50
- Recall : 0.50
- F1 score : 0.50
- Accuracy: 0.50

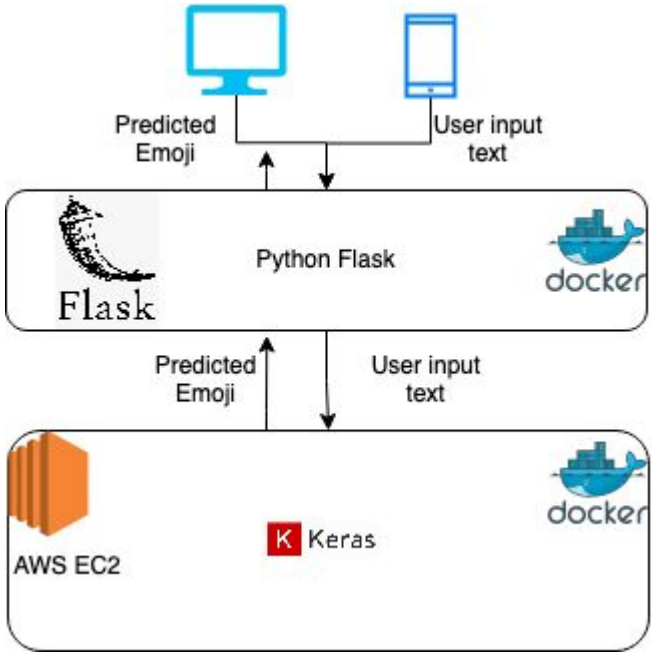


THINK BIG WE DO℠

Deployment

- We developed a simple web interface, which calls a rest API written in python flask. This flask API calls the model predict function from the pre stored weights file.





THINK BIG WE DO™

Try our demo: http://18.141.220.45/

Conclusion and Future work

- Bidirectional LSTM works better than Single layer LSTM and 2 layer stacked LSTM.
- We can try to implement this with CNN model(Conv1D) and see whether CNN performs better than LSTM.
- People have different interpretations of emojis, and often combine multiple emojis without much organization. May be removing the noisy data can improve the accuracy.
- Glove embedding has different types like 100D and 200D, trying with different embedding may improve the accuracy.

Try our demo: http://18.141.220.45/

Thank you

Try our demo: http://18.141.220.45/