GIF-based Communication Using Sentiment Analysis

Anurag Bejju , Sachin Prabhu , Chatana Mandava , Subikshaa Senthilkumar
Department of Computing Science (Big Data)
Simon Fraser University
8888 University Dr, Burnaby, BC V5A 1S6
{ abejju , sthandap , cmandava , ssenthil }@sfu.ca

Abstract

In the modern digital age, *Animated GIF's* have widely become popular on various digital forums, messaging applications and social media platforms. They are gaining momentum in becoming a medium to visually express emotion in today's society. With robust datasets containing GIFs that are emotionally expressive available, state-of-art models are used to prune emotion and detailed description of it. Our project aims to provide a novel approach that takes in text messages from users and identify an appropriate set of GIF's that understand the context and sentiment of it.

1 Introduction

Over the last few years, Animated images – otherwise known as *GIFs* have become a Digital Marketing trend all across the globe. As GIFs are more appealing and effective compared to the usage of images on social media, messaging platforms like Facebook, WhatsApp, Skype and Twitter have started supporting it. With businesses trying to stay relevant with the latest trends around the world, GIFs have become a tool for companies to market their products and services more effectively. With it becoming a vital medium to communicate and express emotions, our project aims to make it easier for people to use GIFs more conveniently in their daily lives.

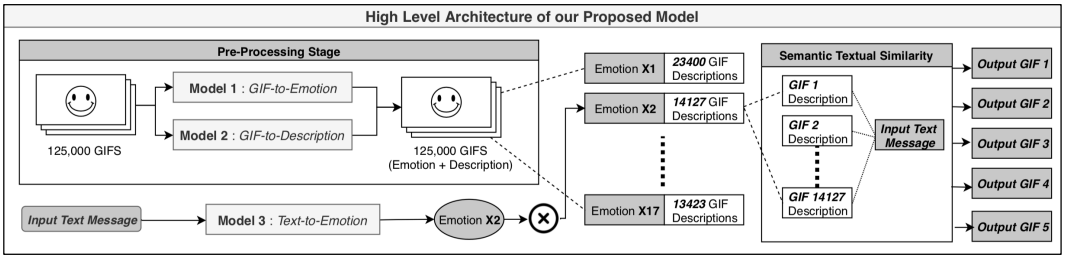


Figure 1: High-Level Architecture of our proposed model

This project proposes a novel approach that takes in text messages from users and identifies an appropriate set of GIF's that understand the context and sentiment of it. As shown in *Figure 1*, in order to construct it, we have used state-of-the-art models that can provide emotion and detailed description of a GIF as well as decipher the overall sentiment of the inputted text message. This was achieved by integrating models that can handle each of the above-mentioned tasks individually. Some of the major challenges that had to be resolved while developing this model are (i) *Lack of robust datasets that provide description and sentiment of GIFs* (ii) *Successful integration of multiple models that can provide a functional mechanism to output a set of contextually relevant GIF's.*

This paper provides a detailed analysis on how we can address these challenges and proposes a unique model that inputs a text message in order to find the best set of GIFs that understand sentiment and context of it. *Section 2* and *3* of this paper, provide a glimpse of related work done in this field and explain about the data-collection methodology used. *Section 4* and *5* provide a detailed methodology of each task that is part of our proposed model as well as share some qualitative results obtained from it. Finally, *Section 6* states the conclusion of this project.

2 Related Work

Historically, traditional sentiment analysis was mainly used to detect emotion from text. Gradually, with the rise in usage of images and GIF's on various social media platforms lead to an increase in research of various image and visual sentiment analysis methods.

Dazhen Lin et al [1] discussed how to do the sentiment detection from a GIF using semantic sequence and proposed a SYNSET forest method to select the SentiPair labels that solves the semantic gap problem. Three important features, namely: *emotional correlation, universality* and *detectability* are considered in this model. Using these three they have constructed a SentiPair. A SentiPair is a joint name of Adjective Noun Pair (ANP) and Verb Noun Pair (VNP). This experiment and the result seemed to be convincing but they failed to predict the accurate SentiPair. The SentiPairs are strongly related to only three set of emotions and dataset containing SentiPairs was very small.

On the other hand, *Morency et al.* [2] proposed a structure which used video sounds and facial expressions to analyze the video clips. They have mainly focused on the sentiment analysis towards video with the limited contents, similar patterns and the average noises. The result was satisfactory but this model cannot deal with large-scale GIF. It is only flexible for small-scale data.

Du Tran et al [3] proposed learning spatiotemporal features with 3D convolutional networks. They tried to prove 3D convolutions are best compared to spatiotemporal features. They have shown that these learned features with simple linear classifier using 3D convolutional network can yield a better performance on various video analysis tasks. But when this is compared with Imagenet baseline the proposed model is still worse because of using a lesser resolution compared to Imagenet and also the model is not trained on all the dataset available.

3 Dataset

One of the major components of our proposed model is the generation of emotion and description from a GIF datatype. In order to achieve this two different datasets (*GIFGIF , TGIF*) were used.

3.1 GIFGIF Dataset

Our goal to extract emotions from the GIFs was formulated using GIFGIF dataset, a crowd-sourcing platform that permits users to vote on animated GIFs based on their emotions. This platform uses GIFs from Giphy that belong to a wide range of categories including movies, sports, TV shows, user-generated content etc. This is essential in order to have a dataset of GIFs with different illumination, effects, resolutions with or without humans or objects in it.

The dataset contains 17 categories of emotions (i.e *amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, happiness, pleasure, pride, relief, sadness, satisfaction, shame* and *surprise*) in total which are based on work by *Paul Ekman's*. The dataset is crowdsourced using the following steps: In the homepage of GIFGIF, a random pair of GIFs are shown to the user and is asked to choose one that best expresses a specified emotion. They are also given an option to choose neither of them. This process is done for all 17 emotions and based on user's votes, each GIF is annotated with scores of it being relevant to an emotion with a higher score denoting higher relevance.

3.2 TGIF Dataset

To extract descriptions from GIFs, Tumblr GIF (TGIF) dataset was used. TGIF dataset contains 100k GIFs with descriptions/sentences annotated by crowdsourcing it. Later, syntactic and semantic validation of the descriptions was done and the dataset was split randomly into training, validation and testing data. Using this, *Yuncheng et al*. designed a model that uses various methods including *Nearest Neighbour (NN), Statistical Machine Translation (SMT)* and *Long Short-Term Memory (LSTM)* to generate descriptions from a GIF. Based on his results, Fine-tuning the LSTM model produced the best results for this case.

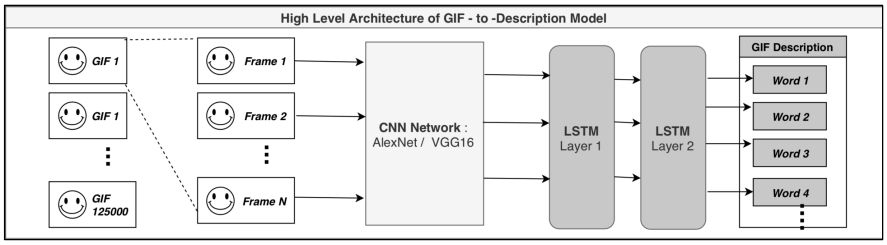


Figure 2: High-Level Architecture of GIF-to-Description Model

As shown in *Figure 2*, the model using LSTM, samples a fixed set of N frames from the GIF and passes it to AlexNet containing a 7 layered CNN with ReLU activation function for non-linearity and faster results. The frames sampled from the GIF are also converted to optical flow images and passed to VGG16, a 16 layered CNN with the weighted average of both being passed on to a stacked LSTM. The first LSTM takes in the output and passes encoded visual features of the frames to another LSTM that encodes the sequence of words obtained from the first LSTM to generate a description of the GIF. This decoding stage begins with <BOS> which indicates the beginning of a sentence. The decoding stages end with <EOS> indicating the end of the sentence. The final dataset containing URL of the GIFs and the corresponding generated description is formed from the predictions made from using the above model.

4 Methodology

Using the data from the above-mentioned datasets, we have constructed a model that takes in a text message to provide a set of GIF's that can understand the context and sentiment of it. As captured in *Figure 3*, The below three sections provide a detailed methodology of each component of our proposed model.

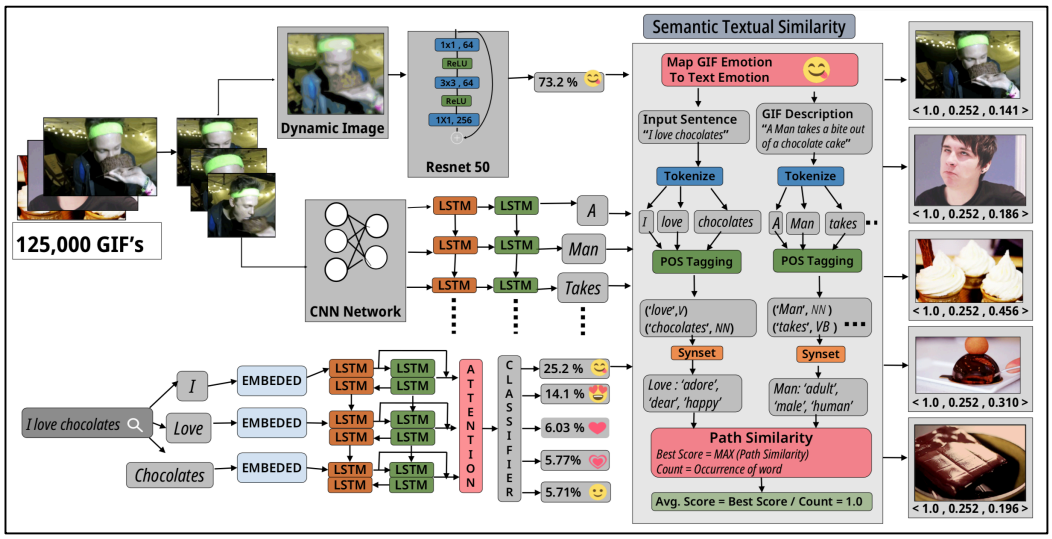


Figure 3: Detailed Architecture of our Proposed Model

4.1 Emotion from GIFs

As detailed in *Figure 4,* To obtain emotions from GIFs, GIFGIF dataset containing scores for 17 categories of emotions and ResNet50 [4], a 50-layer deep convolutional neural network was used. Theoretically, deeper the CNN, the performance of the network should improve but when experimented, this deteriorates the overall performance of the existing network. To solve this problem, residual networks containing shortcuts that allow networks to be deep and still perform better were introduced. These perform better because they work as assembling subsets of residual modules and the use of 3-block residual network has proven to be a great classifier for images.

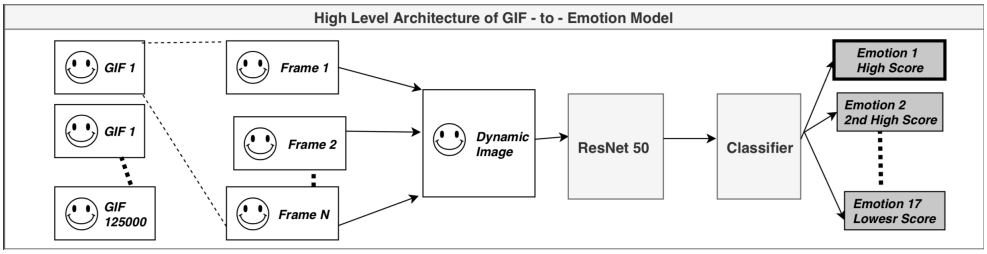


Figure 4: High-Level Architecture of GIF-to-Emotion Model

Since ResNet50 requires the input to be of image type, we convert input GIF to a dynamic image by implementing dynamic image technology mentioned in the work done by Bilen et al.[5] . Frames from a GIF are taken and are converted into a dynamic image which encodes all the transitions involved in frames of a GIF. Dynamic images are obtained by direct rank pooling on the raw image pixels of a GIF generating one RGB image per GIF. This idea is simple but powerful as it enables the use of existing CNN models directly on video data.

The above created image is used as an input feature and the scores for emotion categories as labels to train the ResNet50 model. Since the training dataset containing 4589 GIFs was small, 5-fold cross-validation was performed and for each fold, the model was trained for 50 epochs. The model with the best validation loss at every fold was saved and the overall best model with least validation loss was chosen from the 5 saved models. A softmax function was used at the final layer of the model and a greedy algorithm was used to make the predictions on the test dataset. In a greedy algorithm, class with the highest probability score is chosen to be the class of the input feature (i.e in this case emotions of a GIF). An accuracy of *64.5%* was achieved on the test dataset using the ResNet50 model.

4.2. Text-to-Emotion

Analyzing emotional content is a crucial part of natural language processing. With its applications being enormous, there has been a significant amount of research and many state-of-the-art models developed in this field. As shown in *Figure 5*, The torchMoji / DeepMoji [6] model developed by students at MIT is a perfect fit for constructing the text-to-emotion part of our proposed model. We have decided to use this model as emoticons are a perfect representation for understanding text's emotion as well as the context of it. For example, "*I have missed my bus again, Amazing! >:[*" sentence can be miss-classified as "*happy*" because of the word *amazing* but *angry emoji* used suggests that the sentence was used in a *negative connotation*.

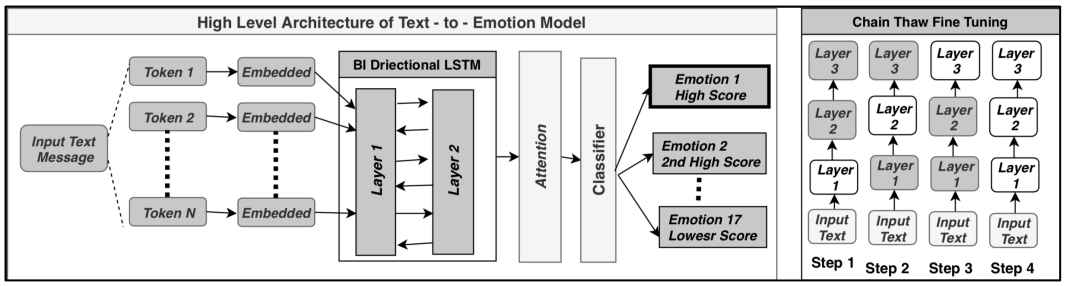


Figure 5 : Architecture of Text- to-Emotion Model & Chain Thaw Fine Tuning Method

TorchMoji model was trained using a huge dataset of 1.6 billion tweets [7]. This model first tokenizes the input sentence and pass it to two *bidirectional-LSTM* layers that can benefit from information present in layers earlier in the network. Since most unidirectional LSTM models learn representations from previous time steps, it becomes a problem when we are required to learn representations from future time steps to better understand the context and eliminate ambiguity.

This can be solved by using a Bidirectional LSTM that have two connections, one going forward in time providing information about previous states and another going backward in time, providing information about future states. This model follows a 'chain-thaw' fine-tuning procedure, which repeatedly unfreezes part of the network and trains it. This process first trains any new layers, then individually fine-tunes each layer from first to the last layer and finally train the complete model in its totality [*Figure 5*]. This is done to avoid overfitting of data for a particular domain and be more adaptive to statements in new domains.

Also in order for the softmax layer to have access to all previous layer, it follows a simple attentional mechanism that takes all prior layers of the LSTM model as input. This softmax layer gives us a probability distribution for each input statement to be a part of the said 17 emotions. We choose the one with the highest probability to map with the GIF's emotion. This pre-trained model had an accuracy of *82.4%*.

4.3. Semantic Textual Similarity

After the input text's emotion is mapped with the emotions of preprocessed GIF dataset, we find the semantic textual similarity between the actual text message and the subset of GIF descriptions portraying the same emotion. In order to do this, *Wordnet* [8] library is used to find the sentence similarity between two sentences and provide an appropriate set of GIF's. This process involves tokenizing of two sentences and applying POS tagging on these tokenized sentences.

POS tagging is a part of speech tagging where you get to know whether the word is noun or verb or adjective etc. The next step is to apply SYNSET function on the output obtained from POS tagging. SYNSET is used to find the synonyms of the tokenized words which will make our work much easier in finding the similarity between two sentences.

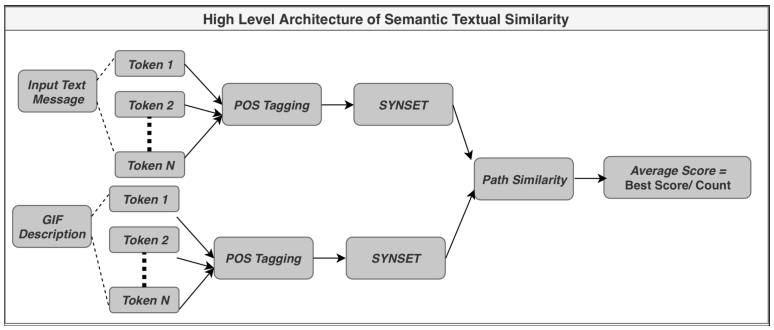


Figure 6: High-Level Architecture of Semantic Textual Similarity

After getting the synonyms a *path_similarity* function is used to generate a score between 0 and 1 depending on the similarity between two words. In this case, for each word in the first sentence, we are checking for the similarity with the words in the second sentence to find the maximum score (i.e best score). Also, the occurrence of each word is considered to calculate the average of the best score and count.

The resultant score is used to rank the subset of GIF's with the same emotion as input text. A GIF which has average score 1 is known as the most appropriate GIF and is ranked first. There are many other traditional methods to find the word similarity like word space model and random indexing but, this model gives more accurate results compared to others.

4 Results

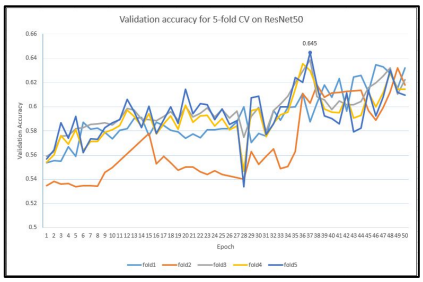


Figure 7: Validation Accuracy for 5 Fold CV on ResNet 50 per Epoch

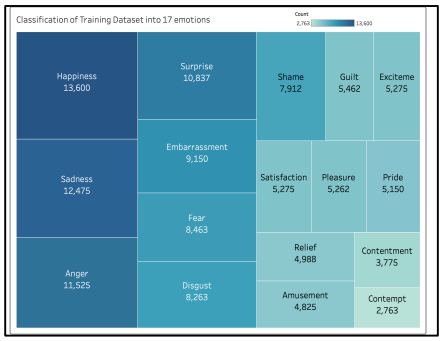


Figure 8: Classification of Training Dataset into 17 Emotions

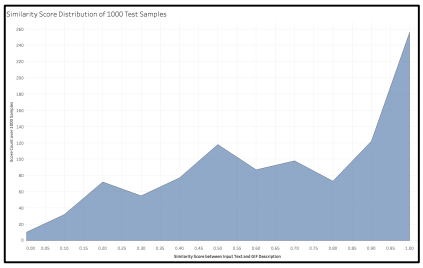


Figure 9: Text Similarity Score Distribution for 1000 Samples

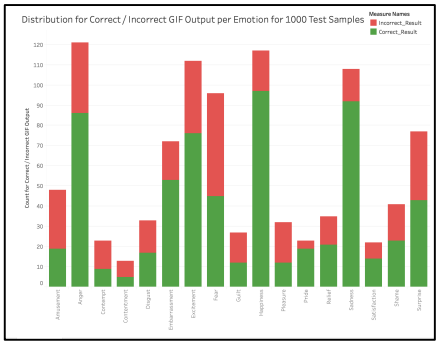


Figure 10: Distribution for Correct / Incorrect GIF Output per Emotion for 1000 Test Samples

To construct our proposed model, we have used several state-of-the-art methods and techniques with varying evaluation metrics. In this section, we will provide details about the performance of each method as well as state some qualitative results of our final model *[Check Appendix 1]*.

Initially, a TGIF dataset containing 125,000 GIFs and their description was compiled using CNN and stacked LSTM models. The same dataset was also converted to dynamic images and ResNet 50 Module was applied to classify that GIF into one of the 17 emotion categories. Since this model was generated by using a small training dataset, 5-fold cross-validation was performed and for each fold, the model was trained for 50 epochs (*Figure. 7*). When the 125000 GIF Dataset was passed in the above-trained model, it resulted in the highest number of GIF's (13,600) being classified as "*Happy*" and lowest number of GIF's (2,763) being classified as "*Contempt*" (*Figure 8*).

Once the preprocessing of GIF dataset is done and each GIF has a description and emotion assigned, we have tested our proposed model with 1000 input text messages. These messages were tokenized and passed to a bi-directional LSTM model with the output passed into a softmax classifier to get the most probable emotion the input text can have. This resulted in the highest number of text messages (121) being classified as "*Angry*" and the lowest number of GIF's (13) being classified as "*Contentment*" (*Figure 10*)

Later, these text messages are mapped with a subset of GIFs having the same emotion and the semantic textual similarity is applied to it. If the description of the GIF and the input text message have the same context, then the similarity score would be 1. If not, the score would range anywhere between 0 to 1. As seen in *Text Similarity Score Distribution for 1000 Samples plot*, 256 text messages had a perfect match (Score = 1) (*Figure 9*). Based on these scores, a set of 5 GIF's is outputted. We have used human validation method to check if the suggested GIF's have the same context and emotion of the input text. Based on our evaluation, we had the best scores for GIFS portraying emotions like *happiness* and *sadness* and least scores for GIFS portraying emotions like *fear* and *amusement*. (Fig 10). Based on this result the overall accuracy of the proposed model was *64.43 %*.

5 Conclusions

Our project involves multiple state-of-the-art deep learning models that can effectively perform various visual and image sentiment analysis tasks. Initially, we have collected the dataset of GIFs with generated descriptions from work done by Yuncheng et al. using a stacked LSTM and CNN models. Later as part of our preprocessing task, emotions were extracted from GIFs. In order to do this, a ResNet50 model was trained with dynamic images of GIFs from GIFGIF dataset to produce results belonging to 17 categories of emotions.

Once the preprocessing of GIF dataset is completed, an input text message is passed into a pre-trained LSTM torchMoji model that uses *chain-thaw transfer* learning method to identify emotions out of it. This is mapped with the emotions of preprocessed GIF dataset, and semantic textual similarity between the actual text message and the subset of GIF Descriptions having the same emotion is found. Top 5 GIFs with highest similarity index produced by WordNet are suggested to the users.

Contributions

Anurag: Researched, analyzed and implemented *Text-to-Emotion* part of this project. I have also consolidated various methods and models implemented by my teammates to construct our proposed model that inputs a text message from a user and identify an appropriate set of GIF's that understand the context and sentiment of it. Finally tested and compiled some qualitative results that measures overall performance of our proposed model.
Subikshaa and *Sachin*: Understanding the implementation and methodology of work by Yuncheng et al. on generating the descriptions from GIFs. Also performed identification of the GIFGIF dataset and trained the ResNet50 model to obtain emotions from the GIFs.
Chatana: Researched various state-of-the-art models that have been used to perform visual and image sentiment analysis. Also worked on semantic textual similarity part of the project.

References

[1] Lin, Dazhen, et al. "*GIF Video Sentiment Detection Using Semantic Sequence*." Mathematical Problems in Engineering (2017).

[2] Morency, Louis-Philippe, Rada Mihalcea, and Payal Doshi. "*Towards multimodal sentiment analysis: Harvesting opinions from the web.*" Proceedings of the 13th international conference on multimodal interfaces. ACM, (2011).

[3] Tran, Du, et al. "*Learning spatiotemporal features with 3d convolutional networks*." Proceedings of the IEEE international conference on computer vision. (2015).

[4] He, Kaiming, et al. "*Deep residual learning for image recognition*." Proceedings of the IEEE conference on computer vision and pattern recognition. (2016).

[5] Bilen, Hakan, et al. "*Dynamic image networks for action recognition*." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016.)

[6] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, Sune Lehmann "*Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm*" Empirical Methods in Natural Language Processing (2017)

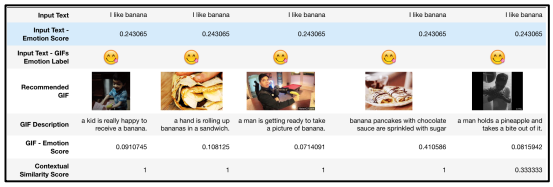
[7] Ji Ho Park, Peng Xu, Pascale Fung et al. "*PlusEmo2Vec at SemEval-2018 Task 1: Exploiting emotion knowledge from emoji and #hashtags*" Association for Computational Linguistics (2018)

[8] Liu, Hongzhe & Wang, Pengfei. "Assessing Sentence Similarity Using WordNet-based Word Similarity". Journal of Software. (2013).

Appendix I

```
Tokenizing using dictionary from /Users/anuragbejju/ML_Project/model/vocabulary.json
Loading model from /Users/anuragbejju/ML_Project/model/pytorch_model.bin.
TorchMoji(
  (embed): Embedding(50000, 256)
  (embed_dropout): Dropout2d(p=0)
  (lstm_0): LSTMHardSigmoid(256, 512, batch_first=True, bidirectional=True)
  (lstm_1): LSTMHardSigmoid(1024, 512, batch_first=True, bidirectional=True)
  (attention_layer): Attention(2304, return attention=False)
  (final_dropout): Dropout(p=0)
  (output_layer): Sequential(
    (0): Linear(in_features=2304, out_features=64, bias=True)
    (1): Softmax()
```

This represents the schema of *TorchMoji* model use to prune emotion out of input texts. Below are some examples used to evaluate our model.

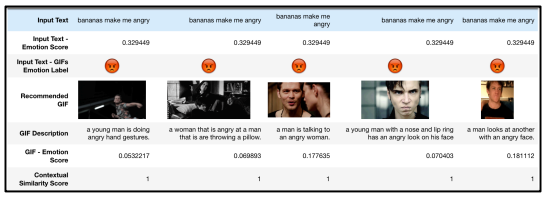


Input Text	I like banana	I like banana	I like banana	I like banana	I like banana
Input Text - Emotion Score	0.243065	0.243065	0.243065	0.243065	0.243065
Input Text - GIFs Emotion Label					
Recommended GIF					
GIF Description	a kid is really happy to receive a banana.	a hand is rolling up bananas in a sandwich.	a man is getting ready to take a picture of banana.	banana pancakes with chocolate sauce are sprinkled with sugar	a man holds a pineapple and takes a bite out of it.
GIF - Emotion Score	0.0910745	0.108125	0.0714091	0.410586	0.0815942
Contextual Similarity Score	1	1	1	1	0.333333

(1) "*I like banana*" returns a set of GIF's that showcase "*happy*" emotion. Although the first 4 GIF's have banana and result in a contextual similarity score of 1, the 5th GIF has a low score (0.333) because he is eating a different fruit.

Input Text	I hate banana	I hate banana	I hate banana	I hate banana	I hate banana
Input Text - Emotion Score	0.415244	0.415244	0.415244	0.415244	0.415244
Input Text - GIFs Emotion Label					
Recommended GIF					
GIF Description	a man is eating a banana and receives something	a young boy is eating banana while watching his friends.	a girl is eating a banana as she sits next to a window in a moving vehicle	a small, black puppy is licking a banana	a young man puts a banana in front of his face and then peels it with his mouth.
GIF - Emotion Score	0.324045	0.130538	0.262053	0.136748	0.214204
Contextual Similarity Score	1	1	1	1	1

(2) "*I hate banana*" returns a set of GIF's that showcase "*hate*" emotion. Also, they portray dislike for banana leading to a contextual similarity score of 1.

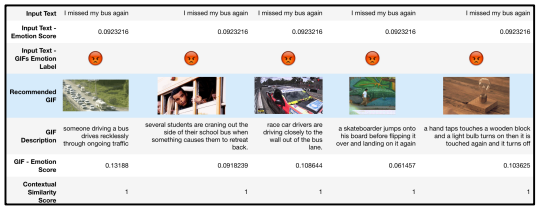


Input Text	bananas make me angry	bananas make me angry	bananas make me angry	bananas make me angry	bananas make me angry
Input Text - Emotion Score	0.329449	0.329449	0.329449	0.329449	0.329449
Input Text - GIFs Emotion Label					
Recommended GIF					
GIF Description	a young man is doing angry hand gestures.	a woman that is angry at a man that is are throwing a pillow.	a man is talking to an angry woman.	a young man with a nose and lip ring has an angry look on his face	a man looks at another with an angry face.
GIF - Emotion Score	0.0532217	0.069893	0.177635	0.070403	0.181112
Contextual Similarity Score	1	1	1	1	1

(3) "*Bananas make me angry*" returns a set of GIF's that showcase "*angry*" emotion. Also, the humans in the GIF's have angry faces leading to a contextual similarity score of 1.

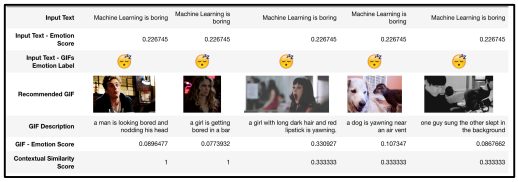
Input Text	I love to party	I love to party	I love to party	I love to party	I love to party
Input Text - Emotion Score	0.210931	0.210931	0.210931	0.210931	0.210931
Input Text - GIFs Emotion Label					
Recommended GIF					
GIF Description	a woman is singing a song whilst wearing a party hat.	men are dancing at a party, one is topless.	a young woman at a party holds a drink and gestures with her hand.	a guy is spinning records in front of a large crowd dance party.	a man and a woman are dancing together at a party.
GIF - Emotion Score	0.0969383	0.0787728	0.384973	0.18003	0.114135
Contextual Similarity Score	1	1	1	1	1

(4) "*I love to party*" returns a set of GIF's that showcase "*enjoy*" emotion. Also, the GIF's descriptions have word "party" leading to a contextual similarity score of 1.



Input Text	I missed my bus again	I missed my bus again	I missed my bus again	I missed my bus again	I missed my bus again
Input Text - Emotion Score	0.0923216	0.0923216	0.0923216	0.0923216	0.0923216
Input Text - GIFs Emotion Label					
Recommended GIF					
GIF Description	someone driving a bus drives recklessly through ongoing traffic	several students are craning out the side of their school bus when something causes them to retreat back.	race car drivers are driving closely to the wall out of the bus lane.	a skateboarder jumps onto his board before flipping it over and landing on it again	a hand taps touches a wooden block and a light bulb turns on then it is touched again and it turns off
GIF - Emotion Score	0.13188	0.0918239	0.108644	0.061457	0.103625
Contextual Similarity Score	1	1	1	1	1

(5) "*I missed my bus again*" returns a set of GIF's that showcase "*angry*" emotion. That said only the first three were contextually a good match for the input text. The last two GIFs were a bad fit and were marked low on accuracy while testing our model.

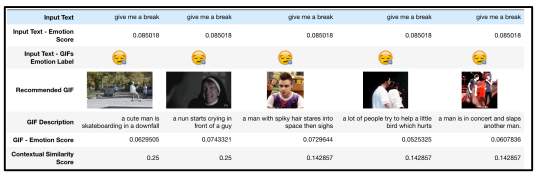


Input Text	Machine Learning is boring	Machine Learning is boring	Machine Learning is boring	Machine Learning is boring	Machine Learning is boring
Input Text - Emotion Score	0.226745	0.226745	0.226745	0.226745	0.226745
Input Text - GIFs Emotion Label					
Recommended GIF					
GIF Description	a man is looking bored and nodding his head	a girl is getting bored in a bar	a girl with long dark hair and red lipstick is yawning.	a dog is yawning near an air vent	one guy sung the other slept in the background
GIF - Emotion Score	0.0896477	0.0773932	0.330927	0.107347	0.0867662
Contextual Similarity Score	1	1	0.333333	0.333333	0.333333

(6) "*Machine Learning is boring*" returns a set of GIF's that showcase "*sleepy*" emotion. The first 2 GIF's descriptions have word bore leading to a contextual similarity score of 1. Other GIF's have a relatively low score as it tries to connect bore to being sleepy

Input Text	I enjoy eating oreos	I enjoy eating oreos	I enjoy eating oreos	I enjoy eating oreos	I enjoy eating oreos
Input Text - Emotion Score	0.482453	0.482453	0.482453	0.482453	0.482453
Input Text - GIFs Emotion Label					
Recommended GIF					
GIF Description	a tray of cookies gets knocked into the air.	a man flips a tray of cookies in the air	a chocolate cookie is being placed in a white ice cream cone	a boy with blond hair is biting into a chocolate chip cookie.	a pop star is feeding pink heart-shaped cookie to somebody.
GIF - Emotion Score	0.0758156	0.129978	0.251603	0.107972	0.0546665
Contextual Similarity Score	0.5	0.5	0.5	0.5	0.5

(7) "*I enjoy eating Oreos*" returns a set of GIF's that showcase "*happy*" emotion. This is a great example to see the contextual connection made between Oreos and cookies.



Input Text	give me a break	give me a break	give me a break	give me a break	give me a break
Input Text - Emotion Score	0.085018	0.085018	0.085018	0.085018	0.085018
Input Text - GIFs Emotion Label					
Recommended GIF					
GIF Description	a cute man is skateboarding in a downhill	a nun starts crying in front of a guy	a man with spiky hair stares into space then sighs	a lot of people try to help a little bird which hurts	a man is in concert and slaps another man.
GIF - Emotion Score	0.0629505	0.0743321	0.0729644	0.0525325	0.0607836
Contextual Similarity Score	0.25	0.25	0.142857	0.142857	0.142857

(8) "*Give me a break*" returns a set of GIF's that showcase "*sad*" emotion. But since there is no GIF data that can properly depict the context, it results in a low contextual similarity score