Emotion Detection in COVID Instagram Posts

Praneeth Chandra, Saurav Chennuri, Keanu Nichols, Leo Seoane, Zilu Tang

(Our code and prediction result is at
https://github.com/PootieT/covid_instagram_sentiment_public)

Introduction

Emotions can be subtly expressed in text. In the current age of social media, users are expressing emotions through text ubiquitously in different platforms. In this project, we looked into ways of extracting emotion from social media posts, and analyze the emotions expressed throughout current events. Specifically, we investigated whether the emotions expressed in Instagram posts during the COVID pandemic are correlated with the presence of Asian people. We used a fine-tuned BERTweet model on Twitter Emotional Intensity Dataset to classify emotions of Instagram post's captions, and verified that there is a statistically significant correlation between emotions expressed in captions and presence of East Asian people. Additionally, we trained a model to detect the presence of an Asian person in the Instagram Post for further validation.

Throughout the analysis of this report, we specifically investigated the following questions:

- How good are emotion classification results on Instagram posts given captions?
- How good are emotion classification results on Instagram posts given facial features in the images?
- How consistent are emotion classifications given a text and image classifier? Can the error be remedied?
- For each of the modality, are there correlations between emotion classification and the presence of east Asian people?

With the above questions, we first introduce the background of emotion in text, image, as well as the anti-asian sentiments during the pandemic. We then describe the dataset we used to train our text and image classifiers. Finally we present the experiments we conducted to answer the questions we described above.

Background

COVID Pandemic

The ongoing COVID-19 pandemic has had a major impact in almost every sector of life across the world. Since the start of the pandemic in December of 2019 in Wuhan, China, feelings about east asian people have taken a negative turn,especially in America. Former President Donald Trump made several mocking remarks towards asian people at this time, and other popular conservative news outlets have spread misleading or false information regarding

the origin of the virus. As a result of negative portrayals of east asian people in media and by powerful political figures, our research team is attempting to monitor and quantify any rise in discrimination through Instagram, where millions of people share their thoughts on current topics through images and text.

Emotions in Text

Emotions decoction, a sub-branch of sentiment analysis in natural language processing (NLP), deals with detecting pre-determined emotion given a textual input. The branch of research is interdisciplinary, from borrowing definitions of emotional models from the field of psychology [Ekman, 1999][Plutchik, 1980][Ortony, et. al. 1990], to using such emotion detectors in various socially impactful applications such as suicide prevention[Yang, et. al. 2012]. Early emotion detection systems use rule based systems as well as keyword / lexicons to classify emotions in text. With the introduction of machine learning, Naive Bayes (NB), word embeddings, and neural networks have been adapted in the usage of emotion detection[Canales and Martínez-Barco, 2014]. Depending on the particular task, and dataset size, best performing models range from NB, Long-Short-Term Memory (LSTM), Convolutional Neural Network (CNN), to Transformer based models [Acheampong et al. 2020]. Given the amount of data we have and relatively straightforward textual input, we decided to use finetuned transformer based models.

Emotions in Images:

Detecting emotions in images is an area of Computer Vision and Machine Learning that is still being researched and with the rise of face masks due to events like COVID-19 it is an area that still has many improvements to be made. There have been some strides in the area of emotion detection with the introduction of [Challenges in] and this resulted in the FER (Facial Expression Recognition) challenge which resulted in a number of different approaches to the challenge with the top teams utilizing CNN's to implement their solution as was seen in [Goodfellow, et. al. 2016]. This challenge was then further expanded upon in FER+ [Barsoum, et. al. 2016] which looked at having multiple labels for images, in this paper they utilized a DCNN (Deep CNN) in order to go about predicting multiple emotions for an image. With the expanse of different architectures for image classification, the model used ranged from VGG, GoogLeNet, ResNet with the paper utilizing a custom VGG network to perform the emotion recognition.

Faces in Images:

The demand for accurate facial recognition and detection has always been high, especially with regards to law enforcement and public safety. The first major test of using computers to automatically detect faces to find criminals in America was the superbowl 2002 [History of face recognition]. Although the result was a failure, the field continued to grow as computing power, open source databases containing cropped images, and machine learning techniques continued to advance. In 2010, Facebook rolled out a facial recognition tool to

automatically detect faces in instagram posts to tag users who may have been featured. The invention of deep convolutional networks and large face datasets created as a result of the increased use of social media and the popularization of the “selfie”, have helped face recognition algorithms to further gain strength and record competitive accuracies. Recently, the UK has tested more advanced surveillance systems using facial recognition at grocery stores, spurring debates over the ethics of using such technology [Big Brother Watch Team].

Multimodal Emotion Detection

In addition to using just single modalities such as text and images, a few works have explored detecting emotion based on voice/ speech, images, text, and other features together. Datasets such as MELD [Poria, et. al. 2018], a multimodal multi-party dataset for emotion recognition in conversations, has been developed for more subtle situations where emotions occur. Our work focuses mainly on emotion detection through single modality mainly because we also wanted to compare and contrast the performance of different modality.

Datasets

Text Datasets:

The Twitter emotional intensity dataset [cite] is created for the SemEval2017 competition of building an emotional intensity classifier. There are 4 categories for emotions for all the tweets: anger, sadness, joy, and fear. Each tweet has a corresponding intensity label, which is obtained from labelers ranking sets of 4 tweets in terms of the emotional intensity. The intensity ranges from 0 to 1 and there are no overlaps between tweets in different emotion sets. We present the statistics of this dataset in table 1.

emotion	anger	fear	joy	sadness	total
train	857	1147	823	786	3613
valid	84	110	79	74	347
test	760	995	714	673	3142

Table 1: Twitter Emotion Dataset Statistics

Since we are building a classifier, we need discretely labeled text data. Since only Tweets that have a high intensity value for each of the emotions are good data for that emotion, we set a threshold (0.5) to filter for high intensity data points within each emotion category, and discard the rest. We perform this for the train, dev as well as test set of the data. After filtering, there are intotal of 1765 sentences in the train, 1610 sentences in the test, and 159 sentences in the valid set.

BERT based models are large and notoriously overfit. In order to build a robust text emotion classifier with all kinds of input, we searched and collected all possible emotion text datasets in

addition to the provided twitter emotional intensity dataset. We collected only publically available datasets, and those that have discrete emotion labels. We avoided any datasets that may require additional data to determine the emotion of the text (ex. Grounded emotions [Liu, et. al. 2017]) When the emotion labels don't match exactly, we convert some of the classes and discard the rest (valence, arousal). After this process, we ended up with 8 additional datasets.

Datasets	size	Label conversion	Description
ISEAR[Scherer and Wallbott 1994]	7667	Remove irrelevant emotions	Obtained from cross-cultural studies in 37 countries and contains 7665 sentences annotated for joy, sadness, fear, anger, guilt, disgust and shame emotions
Cecilia Ovesdotter Alm's Affect data [Alm 2018]	1383	Remove irrelevant emotions	Constructed from Tales and classified into angry, fearful, happy, sad, disgusted and surprised emotions
DailyDialog [Li et. al. 2017]	13118	Merged "disgust" to "anger"	Contains 13118 Dialogues extracted from conversations and annotated for happiness, sadness, anger, disgust, fear, surprise, and others
CrowdFlower [Sentiment analysis in text]	39740	Merged "empty" to "sadness"; Merged "enthusiasm" to "joy"; Merged "love" to "joy"; Merged "fun" to "joy"; Merged "happiness" to "joy"; Merged "relief" to "joy"; Merged "hate" to "anger"; Merged "worry" to "fear";	Constructed from 39,740 tweets and annotated for thirteen13) emotions
Emotion Stimulus [Ghazi et. al. 2015]	2414	Merged "disgust" to "anger"; Merged "shame" to "sadness";	Data developed from FrameNets' annotated data for emotion lexical unit.Contains 1594 emotion-labeled sentences
MELD data [Poria et. al. 2018]	13711	Merged "disgust" to "anger";	Obtained from dialogues and utterances in a Television Show called Friends

EmotionalPush [EmotionX Datasets]	3716		Obtained from dialogue in Facebook messenger chats.
SMILE dataset [Wang et. al. 2016]	1232	Merged “happy” to “joy”; Merged “happy|surprise” to “joy”; Merged “happy|sad” to “joy”; Merged “sad|disgust” to “sadness”; Merged “sad|disgust|angry” to “anger”; Merged “sad|angry” to “anger”; Merged “disgust|angry” to “anger”; Merged “disgust” to “anger”; Merged “angry” to “anger”;	Gathered from tweets about British Museum

Table 2: Text Emotion Datasets[Acheampong et al. 2020]

In addition to training a good classifier that does well on the dev set of twitter emotion dataset, we also need a model which performs well on instagram data. Since we are performing predictions in a new domain, we investigated the shift in the data. Immediately, we noticed that there are different languages other than English present in Instagram posts. We used FastText [Joulin, et. al. 2016a, Joulin, et. al. 2016b] to determine the language and produced a histogram of all languages present in our dataset:

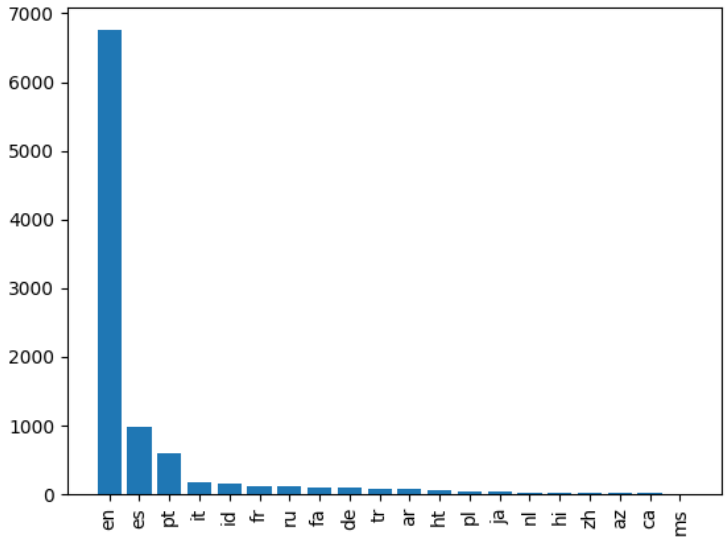


Figure 1: Most frequent languages distributions in Instagram posts

There are in total 67 distinct languages in the dataset, with 30 of them containing more than 5 sentences in the dataset. The most frequent languages are: English (6750), Spanish (979), Portuguese(591), Italian(179), Indonesian(149), French(127), Russian(110), Persian(104), German(97) and Turkish(79).

The goal of our data preprocessing was to translate the non-English sentences to English. But since translational services may contain noise, we want to verify that our model is invariant to the translational noise. So we performed the following experiment. We used the same language frequency found in the Twitter emotion dataset, and created a back-translated version of twitter emotion test set. For each sentence, we randomly translate it to a foriegn language and translate the result back to English. We evaluated the model on the back translated set and found out that the performance of our model decreased due to the translation noise. We realized that emotions may often be lost in translation [Jackson, et. al. 2019] and decided to put less emphasis on foreign posts. For future direction, we can look into multilingual models that are able to learn the emotion for each language in the embedding space without translation.

Test set	Original	Back-Translated
Weighted Avg Accuracy	0.88	0.80

Table 3: accuracy summary of training and testing data on base model. For detailed model description, see Methods-Text Emotion Classification

Face Emotion Detection Dataset:

We utilized the model proposed in [Aaditya1978] which utilized the images in [Challenges in representation learning] which is known as the FER (Face Emotion Recognition) dataset. This data comprises of 48*48 greyscale sized images with the labelled emotions being (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set contained 28,709 examples with the test set containing 3589 examples.

Now for our task of detecting the emotion of the faces from the “Labeled_instagram_posts_related_to_covid” spreadsheet we used columns “Q2_cetegory” is equal to 1 and “Q2A_Type of Human” is equal to 1 and 3 in the spreadsheet which resulted in a dataset of 835 images. After corresponding that with the images present in the “Images.zip” file, this resulted in 406 images in total. We then extracted the faces present in the 406 images which resulted in 395 extracted faces.

Large-scale Face Dataset for Race Detection:

Datasets	size	Label conversion	Description
UTK[Zhang, et. al. 2017]		None	The dataset consists of over 20,000 face images with annotations of age, gender,

			and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc.

One of the biggest challenges in training a race detection model is selecting a suitable dataset. Many facial recognition datasets have disproportionate amounts of white males, and often vary widely in image size, cropping formats, faces obstructed, illumination etc. Because of the variance in instagram pages' content, it is important to represent as many different people and perspectives as possible. The UTK dataset offers a fair amount of diversity in race, age,gender, and image angles. The people in the dataset are 47.7% female, 14.5% east asian, and a wide range of ages. Each image human annotated and labeled with age,date collected,gender, and race in the file name for easy label extraction.

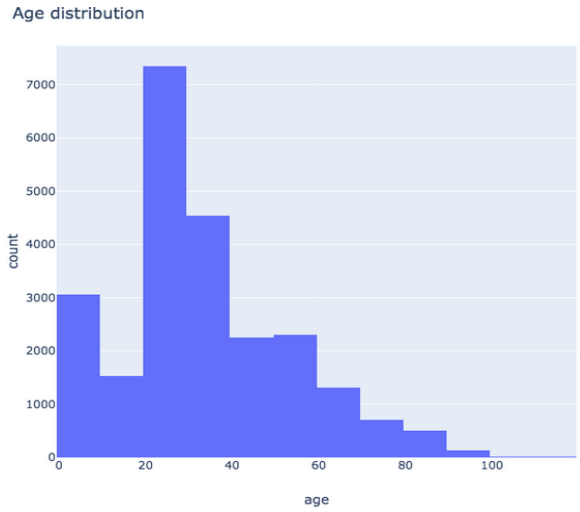


Figure 2
Age Distribution of UTK Dataset

Methods

Text Emotion Classification:

For classifying emotion from text, we fine-tuned a few variants of transformer based models, specifically RoBERTa, BERTweet, and GPT-3. RoBERTa [Liu, et. al. 2019] is a transformer based model trained in a more optimized way than BERT. Compared to its predecessor, RoBERTa is trained with dynamic masking (where each sentence is masked differently 10 times in the dataset), using full sentences that may cross boundaries without next sentence prediction loss , large mini-batches and a larger byte-level BPE [Sennrich, et. al. 2016]. BERTweet, built on top of RoBERTa, uses a special tokenization scheme which includes emoji processing, anonymized Twitter handle and urls. It is also pretrained on large English Twitter datasets to be accustomed to the formats of Tweets. GPT3 is the third iteration of OpenAI's large transformer based model with decoder only module. With the recent release of API endpoints to the public, we wanted to see if the hundred billion parameter model can perform

classification as well as the others. For base models we used: 12 layers, 12 attention heads, 768 embedding dimensions. For large BERT-based models, there are 24 layers, 16 attention head, and 1024 embedding dimensions. On top of BERT-based models, we use the [CLS] token embedding as the input to the classification layer, with two linear layers with 0.1 dropout in between. For the classifier, we output the number of classes in a vector. For the regressor, we output a single scalar. During experiments, the model is trained with ADAM optimizer and a learning rate of 2e-5 for 30 epochs.

Add training detail of GPT3 and post processing methods
For training the GPT3, we need to provide the inputs as a prompt and completion set. Where the completion set is the class we are trying to classify the inputs to. This dataset should be provided as a JSONL file. For the purpose of classification, at the end of each prompt, we need to add a tag “\n\n###\n\n” as a text separator. The model tries to predict the next word based on the input semantically. Using the API calls, we can upload the data and finetune one of the available engines of GPT3. Same goes for the test set, where we upload the jsonl file consisting of a prompt and direct the network to predict.
The output would be a word or a multiple of words usually predicted stochastically. But the predictions usually consist of the emotion word that we are trying to classify the sentence to. Post Processing requires identifying the emotion word in the sequence of predicted tokens. This is where the separator we used in the training prompt comes into use, we can separate the text predicted with the separator tag we used and we can directly filter out the predicted word for the input sentence.

Face Emotion Detection:

To detect the emotion of faces in the images, we first used the opencv face extraction model [Vaessin, et. al. 2021] to extract each of the faces in the Face Emotion Detection Dataset, which resulted in 395 faces. We then input these faces into a CNN that was proposed in [Aaditya1978] which used the FER [Challenges in representation learning] dataset which uses a 48*48 grayscale image and outputs the emotion of that image. The CNN is composed of 4 convolutional layers that each have batch-normalization, ’relu’ activation, max pooling 2D and a dropout layer. These 4 convolution layers are then followed by a flatten layer, which has 2 fully connected layers each with batch normalization, ‘relu’ activation and a dropout layer, and finally we have a softmax activation function with 7 outputs for each of the emotions. The optimizer is Adam with a loss using categorical cross-entropy. The CNN model is trained for 15 epochs with an accuracy of 68% on the validation set.

In order to use this model however we combined some of the labels with “Disgust” and “Fear” being combined with “Angry”, “Neutral” and “Surprise” and “Happy” were labeled as “Joy”. Additionally, this model was chosen since it performed well on the FER dataset and was easy to implement for our task.

Asian Face detection:

For the race detection model, we utilized Rodrigo Bressan's Keras implementation of a basic convolutional neural network to extract low level features from the face, and then added his race detection branch as well to generate probabilities for each class [Rodrigobressan]. The basic CNN layers include 3 convolutional blocks, consisting of a 2d convolutional layer with kernel size (3,3), followed by a relu activation layer, batch normalization, max pooling layer, and then a dropout layer. The race detection layers consist of a dense layer, a relu activation layer, batch normalization and dropout layers. The final two layers are a dense layer followed by a softmax activation layer to transform the input into probability distributions. Cross Entropy Loss was the loss function, with Adam as the learning rate parameter optimizer. The learning rate was initialized at .0004 and trained for 100 epochs. The hyperparameters were selected based on Rodrigo's previous work done on the UTK face dataset. This model was able to achieve an average f1 score of .80 on the test set for race detection across 5 categories: white, black, asian, indian, and others. Next, we used the pretrained model to predict the presence of asian people in our labeled covid images dataset, particularly the presence of asian column, and generated the precision and recall of each class.

Results (Experiments):

Text Emotion Classification:

To verify our model works well for emotion detection, we first used RoBERTa as a regressor to predict emotion intensity, as the original task of Twitter Emotion Dataset asked for [Mohammad and Bravo-Marquez, 2017]. For each sentence in the test set, we predict the intensity and calculate the Pearson correlation between predicted intensity vs true label intensity.

	Anger	Sadness	Joy	Fear	Avg
Prayas (Competition Top Submission) [Mohammad and Bravo-Marquez, 2017]	0.77	0.73	0.76	0.73	0.75
Ours	0.73	0.82	0.81	0.80	0.79

Table 4: Emotion intensity regression comparison

As seen in the table 4, our model beat the highest competition submissions score by a large margin[cite] (thanks to BERT-based models). With such insurance, we are confident the emotion classifier can be equally competitive.

For classification, we trained RoBERTa-base on the twitter emotion filtered train set (**Original**), and 8-dataset combined train set (**Combined**). For comparison simplicity, we just report classification accuracy (we find that it is a relatively good indicator of model performance amongst other metrics, see more tables below. For a detailed classification report, see Appendix Tables 13-15). Note, we accidentally used test set to compare test results, but we will also report the best performing model on dev set for comparison with other teams.

	precision	recall	f1-score	support
anger	0.95	0.93	0.94	43
fear	0.94	0.98	0.96	46
joy	1.00	0.97	0.99	39
sadness	0.90	0.90	0.90	31
accuracy			**0.95**	159
Macro avg	0.95	0.95	0.95	159
Weighted avg	0.95	0.95	0.95	159

Bertweet base results on **dev set**

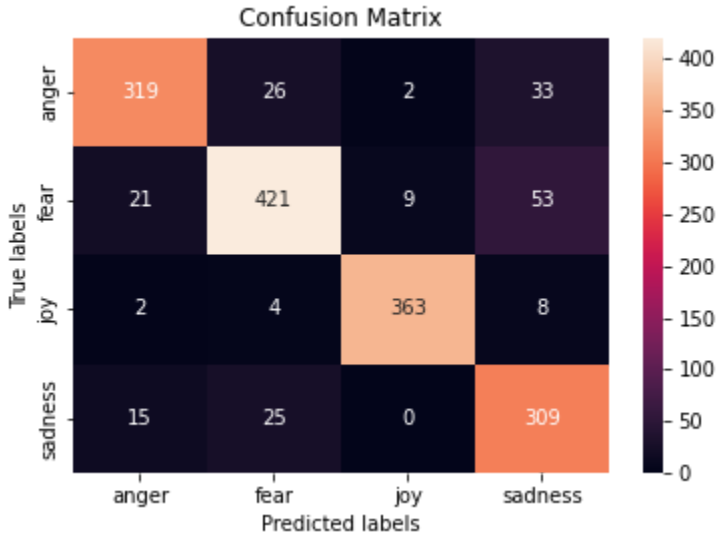
Training dataset	Original	Combined
Accuracy	0.88	0.85

As seen in the table above, the base training set provides enough data to produce a good result for the twitter test set, and we decided to use it for instagram post prediction. The next step is to determine the type of the model to use for classification. Below are the detailed classification reports generated from sklearn.classification_report on the Twitter test set.

	F1 (anger)	F1 (fear)	F1 (joy)	F1 (sadness)	Accuracy (weighted)
RoBERTa	0.85	0.86	0.96	0.82	0.88
BERTweet-base	**0.88**	**0.87**	**0.98**	0.82	**0.89**

BERTweet-large	0.88	0.86	0.97	**0.83**	**0.89**
GPT-3	0.76	0.74	0.89	0.72	0.78

Table 5: Text Emotion Classification Results



As seen from the table above, the BERTweet-base model performed the best across all 4 emotion classes. We used this model and predicted on the instagram post contents. Below is a short summary of the emotion prediction.

Here is an emotion prediction breakdown across the top 5 languages in figure 3.

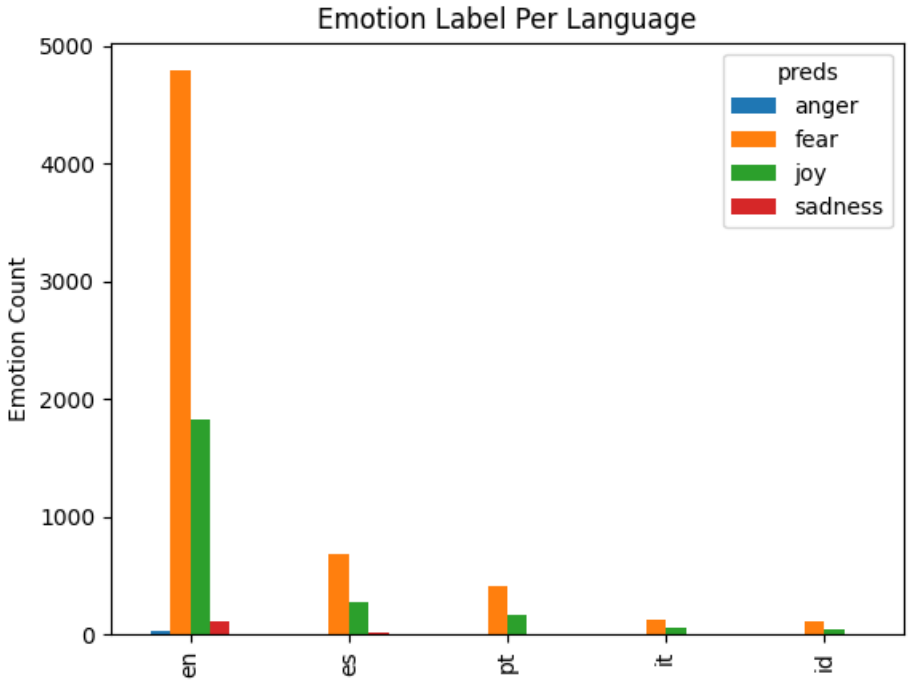
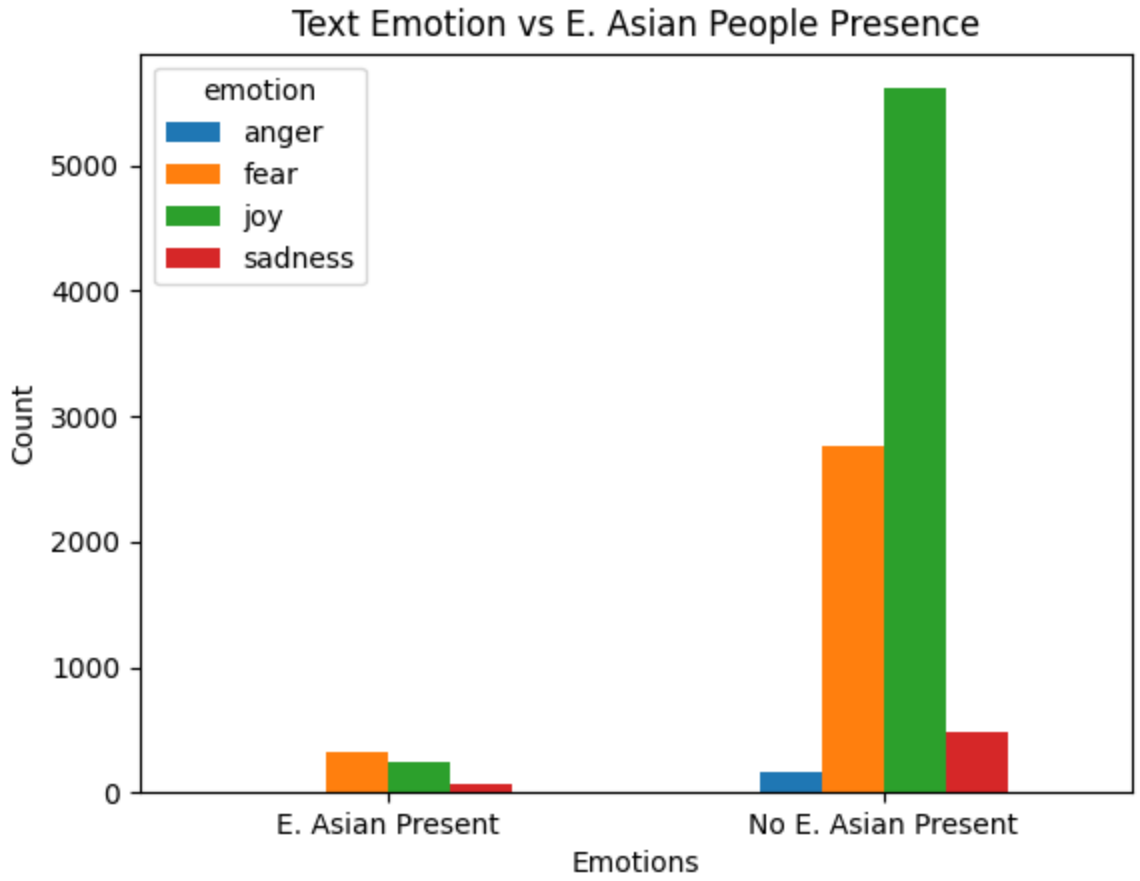


Figure 3: Emotions of Instagram Captions Across top 5 languages

Here is an emotion prediction breakdown given presence of east asian people (5A in instagram dataset feature)

emotion	None East-Asian	East-Asian	Total
Joy	5616 (62%)	236 (37%)	5852 (61%)
Sadness	477 (5%)	60 (9%)	537 (6%)
Anger	163 (2%)	11 (2%)	174 (2%)
Fear	2759 (31%)	326 (52%)	3085 (31%)
Total	9015 (93%)	633 (7%)	9648

Table 6: Contingency table of text emotion classification over east Asian People presence. Chi2 = 160.432, critical=7.815, prob=0.95, degree of freedom = 3, $p < 0.001$, significant



Given the two categorical features (emotion vs. presence of E. Asians), we chose the Chi^2 test as a way to calculate statistical significance. We found that the distribution of emotion is statistically significantly different between those that contain E. Asian and those without.

Face Emotion Classification:

Using the CNN model from [Aaditya1978], we predicted on the Face Emotion Detection Dataset. Here is a breakdown of emotion predictions for all images, we predicted on a subset of the dataset which are the 395 images as described in the Face Emotion Detection Dataset section of the report :

Emotion	Joy	Sadness	Anger	Fear
Count	212	128	52	3

In addition, we provided the contingency table dividing the prediction based on the presence of east Asian people.

emotion	None East-Asian	East-Asian	Total
Joy	115 (56%)	97 (51%)	212 (54%)

Sadness	60 (29%)	68 (36%)	128 (32%)
Anger	28 (14%)	24 (13%)	52 (13%)
Fear	1 (1%)	2 (1%)	3 (1%)
Total	204 (52%)	191 (48%)	395

Table 7: Contingency table of Image emotion classification over east Asian people presence. Chi2 = 2.244, degree of freedom = 3, p = 0.523, not significant / independent

Face vs Text on Emotion Classification:

Since we have two ways of classifying an Instagram post, we discuss the similarities and differences between the two approaches and results. Given that not all of the images contain faces, the number of Instagram posts we are able to use face emotion classifiers on is extremely limited (500 out of 9600). Here we compare the agreements between text and emotion classification:

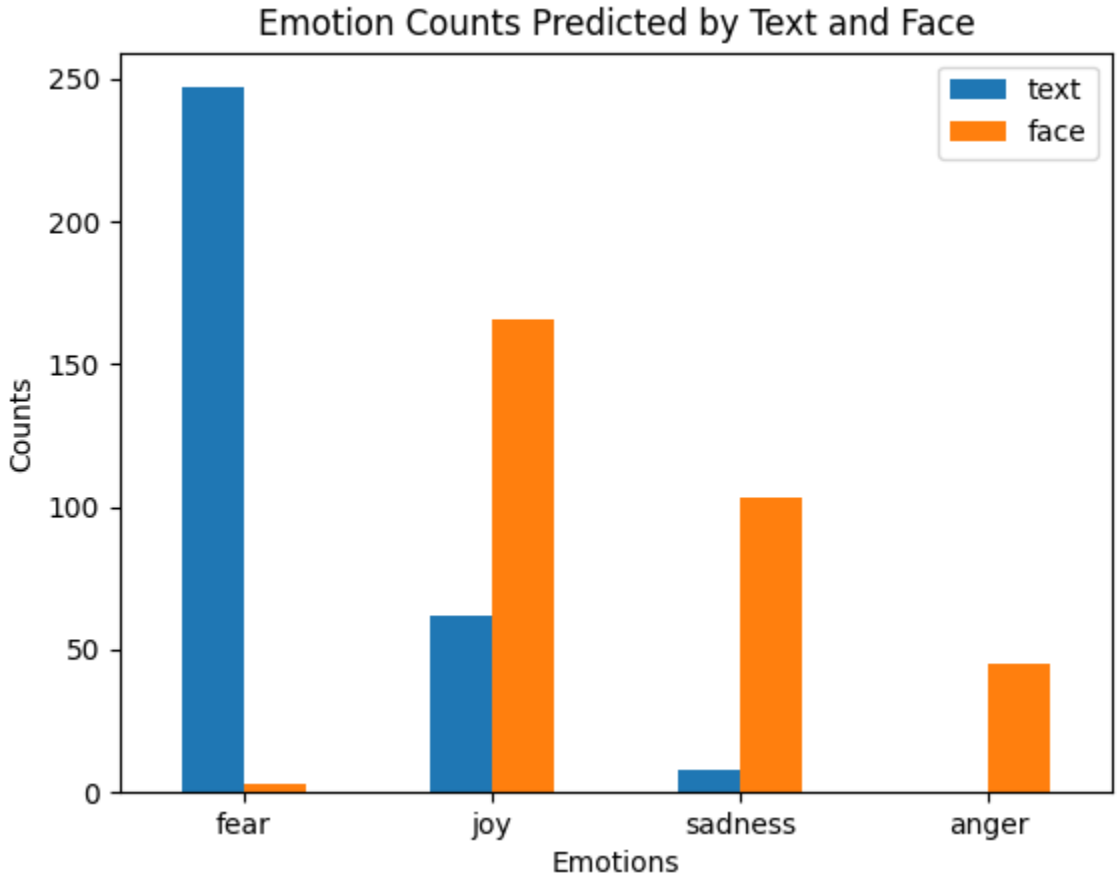


Figure 4: Emotion Classification Distribution with Caption and Face

Emotion	Prediction
Joy	39
Sadness	6
Anger	0
Fear	3

Table 8: Number of predictions in which Text and Facial Emotion Prediction Agree

Emotion	Face Prediction	Post Prediction	Total
Joy	23	127	150
Sadness	2	97	99
Anger	0	45	45
Fear	0	244	244

Table 8: Number of predictions in which Text and Facial Emotion Prediction Disagree

As a naive way of incorporating the results of two modality of prediction together, we use the following rule: if the models agree on the emotion, we use such emotion as the final emotion label. Otherwise, we use emotion given by the model with more confidence. With this heuristic, we classified the subset of instagram posts which contain faces and provided the contingency table again below:

emotion	None East-Asian	East-Asian	Total
Joy	98 (77%)	131 (69%)	229 (72%)
Sadness	23 (18%)	53 (28%)	76 (24%)
Anger	3 (2%)	6 (2%)	9 (3%)
Fear	2 (2%)	1 (1%)	3 (1%)
Total	126 (40%)	191 (60%)	317

Table 10: Contingency table of image and text combined emotion classification over east Asian people presence. Chi2 = 4.805, degree of freedom = 3, p = 0.187, not significant / independent

Asian Face Detection:

UTK Pretrained Model Results:

	precision	recall	f1-score	support
white	0.85	0.89	0.87	3007
black	0.86	0.83	0.85	1361
asian	0.81	0.85	0.83	997
indian	0.72	0.78	0.75	1169
others	0.42	0.22	0.29	506

Table 11: Precision and recall on the test set for UTK-Face

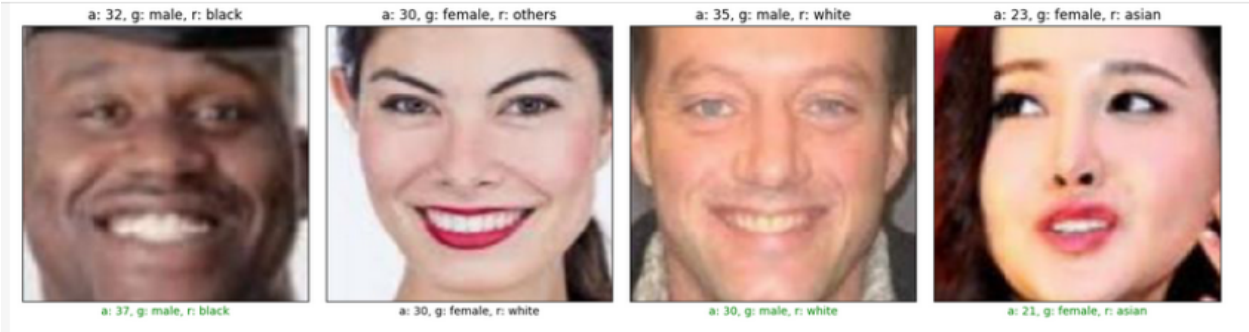


Figure 5 : Results of the model pre-trained on the UTK dataset. The model correctly predicts the gender and race of all but the second image.

Unfortunately, the model was not able to accurately or reliably predict the presence of asian faces in the dataset. We hypothesized that this is due to the faces being covered by masks and camera filters, or stylization and cartoons present in the test set.

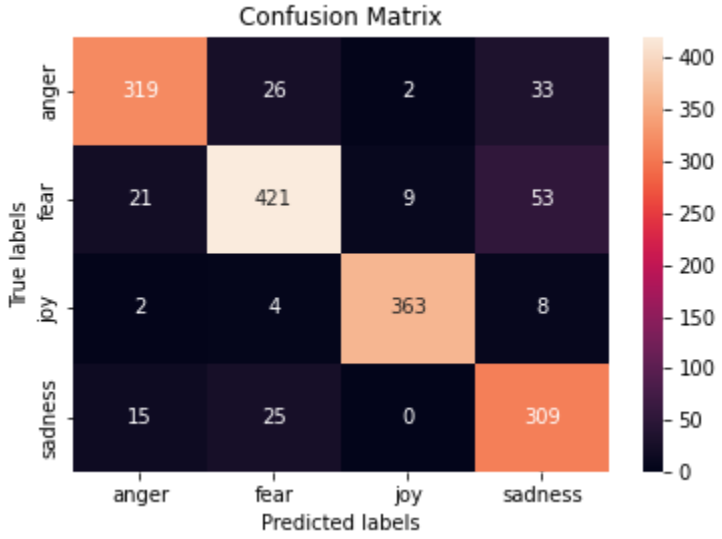
	precision	recall	f1-score	support
Asian(1)	0.50	0.07	0.12	193
Not Asian(2)	0.60	0.95	0.74	5287
accuracy			0.60	480
Macro avg	0.55	0.51	0.43	480
Weighted avg	0.56	0.60	0.49	480

Table 12: Classification report for the true race values vs predicted race
1 denotes the subject is not asian, while 2 means an asian is present in the image

Discussion:

Text Emotion Classification:

BERTWeet Performance: As expected BERTweet performed the best out of all the other models by a small margin. This indicates that by pretraining on a large corpus of Twitter data, BERTweet is able to learn the distribution of words used in Tweets better than baseline models. It is interesting that despite having one of the lowest number of examples per class, joy was one of the easiest emotions to detect. Perhaps it is because “joy” is the only positive emotion out of the four emotions to predict. Other emotions, especially “sad” and “fear”, are much closer in semantic space.



GPT Performance: The low performance of GPT-3 is unexpected. We suspect that re-casting the task of next word prediction to classification from pretrain to fine-tuning is still suboptimal. We noticed that during test time prediction, GPT-3 often output words that are not emotion labels such as “#angry”, “#Trump”, and sometimes we come across synonyms of emotions like “rage” for “angry” and “happy” for “joy” etc. The fact that the fine tuned GPT-3 model is still predicting over all word tokens creates a much harder task for the model, compared to regular BERT based classifiers.

Universal Emotion Across Languages: Despite a small performance drop in emotion prediction across translated posts, we observe similar emotion patterns across different languages. It is only natural that most of the posts on Instagram during COVID-19 are showing fear and some joy. Afterall, such a global scale pandemic is unforeseen and no one is prepared for it. Interestingly, we did not see a lot of sadness in the posts. One rationale is that we are missing a lot of “sadness” posts because they are misclassified as it is often confused with

“fear”. Another, more human explanation is that people do not often post sad posts on social media. In the social platforms, people rant, applaud, but rarely expose their vulnerabilities and announce to the world about their sadness. It would be an interesting comparison to see whether such patterns exist pre-pandemic. Perhaps the occurrence of sadness in real life encourages social platforms to be a place where people share more happy moments of life, and forget about real life for a bit.

Future Steps: During the dataset selection process, we could have spent a bit more time understanding the reason behind degraded performance in aggregated dataset. Given more time, we want to analyze the difference between different text emotion datasets, the distribution and the shift, as well as the differences between more fine-grained emotions such as disgust, love, and sad-happiness. For model selection, we also used the most common large pretrained model approach. However, given the gap from perfect prediction, we want to understand what are the cases we fail to predict on and perhaps try negative sample mining techniques [Robinson, et. al. 2020] to separate emotion classes that are close in embedding space. Additionally, since we noticed the drop in performance due to translational noise, an interesting direction is to look into emotion preserving translation or using multilingual models for sentiment analysis.

Face Emotion Detection:

Error Analysis: The 406 images that were extracted from the “Images.zip” file, after looking through the dataset, even with a human annotator, it was quite hard to see some of the faces in the dataset as seen in **Figure 6a (Top left)**. hence we saw that the extracted faces were only 395 extracted faces.
The 395 extracted faces it was observed that some of the images were mislabeled as faces as seen in **Figure 6b (Top right)**. In terms of the facial emotion detection, we can see how the model at times can detect the emotion of a person as seen for **Figure 6d (Bottom left),** however it can also misclassify an example, for instance for **Figure 6c (Bottom right)**, which was misclassified as sadness

Figures above: In clockwise direction 6a, 6b, 6c, 6d

Figure 6: (Top left) Example of an image where it is difficult to see the faces in the image and by extension the expressions of the people in the image. (Top right) Example of an image that was mistakenly extracted as a face. (Bottom left) Example of an extracted face that was predicted to be the “Joy” label. (Bottom right) Example of an extracted face that was misclassified as “Sadness” when it should have been predicted as “Joy”

Future Directions: In addition to the base model we used in the paper, we wanted to try an additional more sophisticated model that may be able to overcome the “mask” problem (as well as other noises common in Instagram images such as cartoonish figures and filters). A perfect example training regime that might have mitigated this issue is masked image training [Zhao, et. al. 2021][Dosovitskiy, et. al. 2020] In addition, we can extend the emotion prediction to beyond faces. Many images without faces contain plenty of emotion, if we extend the range of image classifiers, we can apply emotion prediction to a large subset of data. Lastly, to deal with image posts where the sentiment is mainly determined through the text in image (i.e. memes), we could also extract the text and supply such information to the text classifier.

Text vs Image Emotion classification

Agreement and Disagreement: We found quite a large discrepancy in the results of the text and image classifier. Most of the times that the models are in agreement occur when people express “joy”.Interestingly, we also observed that the image model has high precision when predicting “sadness” and “fear” (given the small sample size). Assuming both models predicted the text or image correctly, we attribute many of the discrepancies to the way people use the platform. Typically on instagram and other social media platforms, users prefer to use images to convey positive emotions, like posting a happy picture with your friends at dinner. Users usually use the platforms to catalogue and share positive moments in their lives through images, rather than show themselves as afraid or enraged. However, this positive proclivity does not carry over into textual data. Many texts that were labeled as “fear” or “sadness” could have seemingly harmless or innocent pictures, with tirades or damaging rants in the comment section. Further, many users could post an image that appears to be positive, but is in fact a satirical post as revealed in the comments.

Future Steps: An obvious next step is to use multimodal classifiers that take image as well as caption into the classification decision. Often there could be subtle interactions between images and captions that can boost confidence. Even though text is a good baseline, there is much information lost by not looking at the image. Afterall, when humans scroll through Instagram, most of time (though just fractions of a second often) are just spent on the image. However, one can argue that the author of the post puts more effort in crafting the right caption. Lastly, an interesting direction is to analyze the emotion of these text over time. How do the general emotion change over time in different languages? As vaccines are discovered, are there more proportions of “Joy” posts than “sadness” post? Do the shift more with global events or more local events given the language and the location of the post?

Emotion Correlation with Asian Faces:

We found a statistically significant correlation between Instagram captions with the presence of east Asian people in the image. However, such correlation cannot say much without further testing. Posts with east Asian people’s faces can be written by a user who is not an east Asian, in which case the post can be sympathizing or expressing “Asian hate”. On the other hand, posts with east Asian people’s faces can also be written by an east Asian person, in this case, the posts are more likely to describe events they faced or someone they know faced, dominated with emotions other than “anger”. In order to determine which is the case in reality, another attribute that would be interesting to collect is the racial profile of the user who posted the image. Given all such consideration, we will still miss the type of posts that did not mention such “Asian hate” at all but express subtle racially charged sentiments in the text. In such cases, we need a more sophisticated model, trained on toxicity or other dataset to detect such hate speech.

General Face Dataset Discussion:

Difficulties with Instagram Data: During data processing for Instagram data, we ran into several issues when dealing with the labeled covid images excel file and the imaze zip file. The label column names are very vague, and have already been value encoded, so we had to manually figure out which labels correspond to which category. Not all of the post ids correspond correctly to the image prefixes, and some images do not have post ids available. This made it very difficult to match which images in the zip file were also present in the excel. Out of the 40,000 images in the image file, and 10,000 rows in the excel, we were only able to positively match 1200. Out of these 1200, only 500 images had a column for ' is_asian'. This greatly reduces the power of our analysis, and the reported accuracy may differ drastically from the true accuracy. We would also consider using other hyperparameters to improve model performance. Unfortunately, as GPU time was limited, the model was only trained once.

Future Steps: For further analysis, we would be interested in predicting the age and gender of the people in the dataset using the pretrained model to see if there is discrimination occuring more often in certain age ranges, or if it applies more to one gender or another. It is unclear how the face detection model generalizes on cartoonish, or camera filtered images which are common in instagram posts. Because of the nature of social media, anti asian sentiment may be present in popular meme formats, which are often stylized. A labeled dataset containing cartoon vs real categories would be helpful for this task. Also, because the race detection model is not a binary classifier, but includes 4 other major races, we could apply our sentiment analysis in other scenarios, i.e racism on instagram pertaining to hispanic people after immigration spikes.

Conclusion:

The Covid pandemic is entering its second full year, and new variants like Omicron and Delta threaten to keep us in a state of perpetual alert and precautions. Through our work, we were able to create a snapshot in time regarding the many sentiments and emotions surrounding this tumultuous time. Unfortunately, the snapshot we are recording is not positive. We saw higher proportions of fear and anger among Instagram posts with the presence of East asian faces compared to those without. Hopefully, we are able to use the tools and knowledge gained from our research, as well as the research of hundreds of other dedicated scientists to help stop discrimination, and continue educating individuals and increasing love and understanding.

Citations:

Aaditya1978. (n.d.). AADITYA1978/face_expression_prediction: This is a jupyter notebook for recognizing live facial expressions. GitHub. Retrieved December 10, 2021, from https://github.com/Aaditya1978/Face_Expression_Prediction.

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2(7). https://doi.org/10.1002/eng2.12189

Alm ECO. Affect in* Text and Speech. Citeseer. Urbana: University of Illinois at Urbana-Champaign; 2008.

Barsoum, Emad, et al. "Training deep networks for facial expression recognition with crowd-sourced label distribution." Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016.

Big Brother Watch. Co-op doubles 'Orwellian' facial recognition cameras in supermarkets — Big Brother Watch

Bressan, Rodrigo.rodrigobressan/keras-multi-output-model-utk-face: Exploring Keras functional API to build a multi-output model to predict race, gender and sex on UTK Face dataset.

Canales L, Martínez-Barco P. Emotion detection from text: a survey. Paper presented at: Proceedings of the Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days; 2014:37-–43; ACM.

Challenges in representation learning: Facial expression recognition challenge. Kaggle. (n.d.). Retrieved December 10, 2021, from https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

Ekman P. Basic emotions. Handbook Cognit Emot. 1999;98(45-60):16.

EmotionX Datasets. EmotionX 2019. (n.d.). Retrieved December 10, 2021, from https://sites.google.com/view/emotionx2019/datasets.

Ghazi D, Inkpen D, Szpakowicz S. Detecting emotion stimuli in emotion-bearing sentences. Paper presented at: Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics; 2015:152-165; Springer.

Goodfellow, Ian J., et al. "Challenges in representation learning: A report on three machine learning contests." International conference on neural information processing. Springer, Berlin, Heidelberg, 2013.

History Of Face Recognition. Facefirst. History of Face Recognition & Facial recognition software

Jackson, JC , Watts J, Henry TR, List JM, Forkel R., Mucha, P.J, Greenhill S.J., Gray R.D., and Lindquist, K.A. "Emotion semantics show both cultural variation and universal structure". Science 366, no.6472 (2019): 1517-1522.

Joulin, Armand, Edouard, Grave, Piotr, Bojanowski, and Tomas, Mikolov. "Bag of Tricks for Efficient Text Classification".arXiv preprint arXiv:1607.01759 (2016).

Joulin, Armand, Edouard, Grave, Piotr, Bojanowski, Matthijs, Douze, Hérve, Jégou, and Tomas, Mikolov. "FastText.zip: Compressing text classification models".arXiv preprint arXiv:1612.03651 (2016).

Li Y, Su H, Shen X, Li W, Cao Z, Niu S. Dailydialog: a manually labelled multi-turn dialogue dataset; 2017. arXiv preprint arXiv:1710.03957.

Liu V, Banea C, Mihalcea R. Grounded emotions. Paper presented at: Proceedings of the 2017 7th International Conference on Affective Computing and Intelligent Interaction; 2017:477-483; IEEE.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

Mohammad, Saif M., and Felipe Bravo-Marquez. "WASSA-2017 shared task on emotion intensity." arXiv preprint arXiv:1708.03700 (2017).

Ortony A, Clore GL, Collins A. The Cognitive Structure of Emotions. Cambridge, MA: Cambridge University Press; 1990.

Plutchik R. A general psychoevolutionary theory of emotion. Amsterdam, Netherlands: Elsevier; 1980 (pp. 3–33).

Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: a multimodal multi-party dataset for emotion recognition in conversations; 2018. arXiv preprint arXiv:1810.02508.

Robinson, Joshua, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. "Contrastive learning with hard negative samples." arXiv preprint arXiv:2010.04592 (2020).

Rodrigobressan. (n.d.). Rodrigobressan/Keras-multi-output-model-utk-face: Exploring keras functional API to build a multi-output model to predict race, gender and sex on UTK face dataset. GitHub. Retrieved December 10, 2021, from https://github.com/rodrigobressan/keras-multi-output-model-utk-face.

Saravia, Yi-Shin. "CARER: Contextualized Affect Representations for Emotion Recognition." . In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3687–3697). Association for Computational Linguistics, 2018.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Association for Computational Linguistics (ACL), pages 1715–1725

Sentiment analysis in text - dataset by Crowdflower. data.world. (2016, November 21). Retrieved December 10, 2021, from https://data.world/crowdflower/sentiment-analysis-in-text.

Scherer KR, Wallbott HG. Evidence for universality and cultural variation of differential emotion response patterning. J Pers Soc Psychol. 1994;66(2):310.

UTK Face Dataset UTKFace | Large Scale Face Dataset

Vaessin, H., Jesudas, kacha, B., Rosebrock, A., Futami, T., Shah, R., Ayush, Hoffman, D., Huseyn, Falahgs, Nika, Mohamed, Soberon, J. H., Gary, Cyprian, Bloch, Y., Horelvis, Hossein, Nico, … Loyange, T. (2021, July 4). OpenCV face recognition. PyImageSearch. Retrieved December 10, 2021, from https://www.pyimagesearch.com/2018/09/24/opencv-face-recognition/.

Wang B, Liakata M, Zubiaga A, Procter R, Jensen E. Smile: Twitter emotion classification using domain adaptation. Paper presented at: Proceedings of the 25th International Joint Conference on Artificial Intelligence; 2016:15; AAAI.

Yang Hui, Willis Alistair, De Roeck Anne, Nuseibeh Bashar. A Hybrid Model for Automatic Emotion Recognition in Suicide Notes. Biomedical Informatics Insights. 2012;5(1):BII.S8948. http://dx.doi.org/10.4137/bii.s8948.

Zhang, ZF, Song, Y, and Qi, HR. "Age Progression/Regression by Conditional Adversarial Autoencoder." . In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).2017.

Zhao, Yucheng, Guangting Wang, Chong Luo, Wenjun Zeng, and Zheng-Jun Zha. "Self-supervised visual representations learning by contrastive mask prediction." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10160-10169. 2021.

Appendix:

	precision	recall	f1-score	support
anger	0.86	0.79	0.82	380
fear	0.85	0.80	0.82	504
joy	0.91	0.97	0.93	377
sadness	0.76	0.85	0.80	349
accuracy			0.85	1610
Macro avg	0.85	0.85	0.85	1610
Weighted avg	0.85	0.85	0.85	1610

Table 13: RoBERTa emotion classification result trained with additional data

	precision	recall	f1-score	support
anger	0.80	0.73	0.76	380
fear	0.78	0.79	0.79	504
joy	0.87	0.95	0.91	377
sadness	0.77	0.74	0.80	349
accuracy			0.80	1610
Macro avg	0.80	0.80	0.80	1610
Weighted avg	0.80	0.80	0.80	1610

Table 14: RoBERTa result trained with Twitter data predict on back-translated test set

	precision	recall	f1-score	support
anger	0.93	0.80	0.85	380

fear	0.81	0.92	0.86	504
joy	0.95	0.97	0.96	377
sadness	0.85	0.79	0.82	349
accuracy			0.88	1610
Macro avg	0.88	0.87	0.88	1610
Weighted avg	0.88	0.88	0.88	1610

Table 15: RoBERTa emotion classification result trained with twitter data

	precision	recall	f1-score	support
anger	0.91	0.84	0.88	380
fear	0.88	0.86	0.87	504
joy	0.98	0.98	0.98	377
sadness	0.78	0.87	0.82	349
accuracy			0.89	1610
Macro avg	0.89	0.89	0.89	1610
Weighted avg	0.89	0.89	0.89	1610

Table 16: BERTweet-base emotion classification result trained with twitter data

	precision	recall	f1-score	support
anger	0.88	0.87	0.88	380
fear	0.86	0.87	0.86	504
joy	0.97	0.97	0.97	377
sadness	0.83	0.84	0.83	349
accuracy			0.89	1610
Macro avg	0.89	0.89	0.89	1610
Weighted avg	0.89	0.89	0.89	1610

Table 17: BERTweet-large emotion classification result trained with twitter data

	precision	recall	f1-score	support
anger	0.72	0.79	0.76	350
fear	0.84	0.67	0.74	454
joy	0.84	0.95	0.89	353
sadness	0.70	0.74	0.72	313
accuracy			0.78	1470
Macro avg	0.78	0.79	0.78	1470
Weighted avg	0.78	0.78	0.78	1470

Table 18: GPT-3 emotion classification result trained with twitter data