Cross-Cultural Differences of Sentiment in Social Media Posts Related to COVID-19

Hongxuan Chen (hc3275)
Advisor: Prof. John R. Kender
Columbia University

Abstract

Different cultures may react and respond differently given a crisis that has worldwide impact. People in different nations have their own views and emotive response based on their culture, ideas, habits and customs. The governments take different actions to the same event based on their national conditions and public opinion as well. In a same topic, while some may express satisfaction, others might show resentment or other complex emotion. Coronavirus (COVID-19) brought a mix of emotions from different nations caused by the information about and appraisal of the pandemic and decisions taken by their respective governments. Social media was bombarded with posts containing a variety of sentiments on the COVID-19, pandemic, vaccine, and hashtags in the past 2 years. The purpose of this research is to analyze reaction of citizens from the U.S. and China to the COVID-19 and people's sentiment about subsequent actions taken by their governments.

1. INTRODUCTION

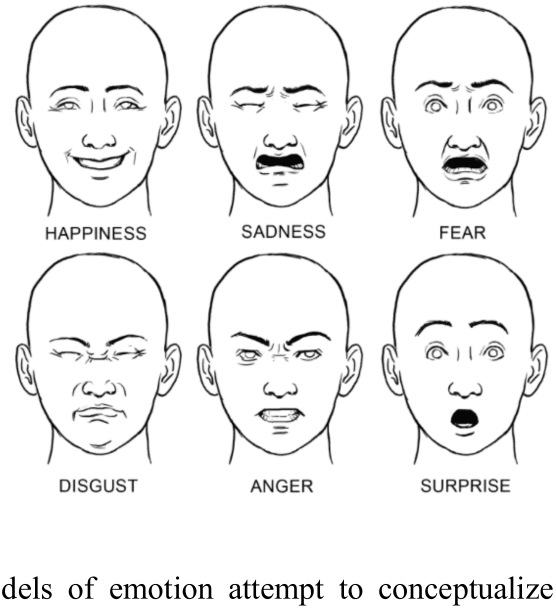
Crises such as natural disasters, global pandemics, and social unrest continuously threaten our world and emotionally affect millions of people worldwide in distinct ways. Understanding emotions that people express during large-scale crises helps inform policy makers and first responders about the emotional states of the population as well as provide emotional support to those who need such support.

The world is seeing a paradigm shift the way we conduct our daily activities amidst ongoing coronavirus (COVID-19) pandemic - be it online learning, the way we socialize, interact, conduct businesses or do shopping. Such global catastrophes have a direct effect on our social life; however, not all cultures react and respond in the same way given a crisis. Even under normal circumstances, research suggests that people across different cultures reason differently. For instance, Nisbett in his book "The geography of thought: How Asians and Westerners think differently and why" stated that the East Asians think on the basis of their experience dialectically and holistically, while Westerners think logically, abstractly, and analytically. This cultural behavior and attitude are mostly governed by many factors, including the socio-economic situation of a country, faith and belief system, and lifestyle. In fact, the COVID-19 crisis showed greater cultural differences between countries that seem alike with respect to language, shared history and culture.

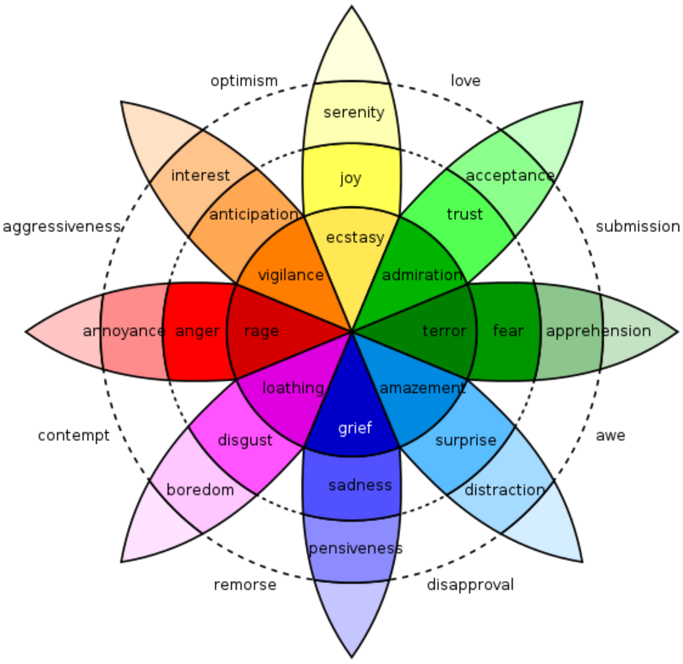
Social media platforms play a vital role during extreme crises, as individuals use these communication channels to share ideas, opinions and reactions with others in order to respond to and respond to crises. Therefore, in this research, we will focus on exploring collective responses to events expressed in social media. Special emphasis will be placed on analyzing people's reactions to the COVID-19 pandemic expressed in the Twitter and Weibo, from which we get data from American and Chinese respectively. Moreover, Twitter is widely popular in the U.S. while Weibo is widely popular in China and both are easy to access with APIs. To this end, the tweets and weibos collected from thousands of users within four weeks of the corona crisis were analyzed to understand how different cultures were reacting and responding to coronavirus. To analyze the data, emotion is an essential aspect to focus on.

Emotion classification, the means by which one may distinguish or contrast one emotion from another, is a contested issue in emotion research and in affective science. Researchers have approached the classification of emotions from one of two fundamental viewpoints that emotions – i) are discrete and fundamentally different constructs and ii) can be characterized on a dimensional basis in groupings.

In discrete emotion theory, all humans are thought to have an innate set of basic emotions that are cross-culturally recognizable. These basic emotions are described as "discrete" because they are believed to be distinguishable by an individual's facial expression and biological processes. Theorists have conducted studies to determine which emotions are basic. A popular example is Paul Ekman and his colleagues' cross-cultural study of 1992, in which they concluded that the six basic emotions are anger, disgust, fear, happiness, sadness, and surprise. Ekman explains that there are particular characteristics attached to each of these emotions, allowing them to be expressed in varying degrees. Each emotion acts as a discrete category rather than an individual emotional state.



Dimensional models of emotion attempt to conceptualize human emotions by defining where they lie in two or three dimensions. Most dimensional models incorporate valence and arousal or intensity dimensions. Dimensional models of emotion suggest that a common and interconnected neurophysiological system is responsible for all affective states. These models contrast theories of basic emotion, which propose that different emotions arise from separate neural systems. Several dimensional models of emotion have been developed, though there are just a few that remain as the dominant models currently accepted by most. Plutchik's model is the most renowned three-dimensional hybrid of both basic and complex categories in which emotions with varying intensities can be combined to form emotional dyads.



The primary objective of this study is to understand how different cultures behave and react given a global crisis. Leveraging some arithmetic strategy, we detect and analyze the emotions of social media posts that is in a same time period and under a same specific topic. Based on the character of the COVID-19 event, we use 6 dominant 'discrete' emotions in this study – positive, sad, angry, scared, surprised, neutral – to label each social media post. In order to detect and classify the fine-grained emotions of posts automatically, we propose a transfer learning method that implement fine-tune training based on some pre-trained models.

2. RELATED WORK

2.1 Responses to Events in Social Media

There is a lot of literature concerning people's reactions to events expressed in social media, which generally can be distinguished by the type of the event the response is related to and by the aim of the study. Types of events cover natural disasters, health-related events, criminal and terrorist events, protests, etc. A recent emerging field of sentiment analysis and affective computing deals with exploiting social media data to capture public opinion about political movements, response to marketing campaigns and many other social events. Studies have been conducted for various purposes including examining the spreading pattern information on Twitter on Ebola and on coronavirus outbreak, tracking and understanding public reaction during pandemics on twitter, investigating insights that Global Health can draw from social media, conducting content and sentiment analysis of tweets.

2.2 Emotion Classification

Hasan et al. utilized the circumplex model that characterizes affective experience along two dimensions: valence and arousal for detecting emotions in Twitter messages. The authors use the emotion words from LIWC3 (Linguistic Inquiry & Word Count) to build the lexicon dictionary of emotions. They extracted uni-grams, emoticons, negations and punctuation as features to train conventional supervised machine learning classifiers and achieved an accuracy of 90% on tweets dataset.

Fung et al. examines people's reaction to the Ebola outbreak on a random sample of tweets. The results showed that many people expressed negative emotions, anxiety, anger, which were higher than those expressed for influenza. The findings also suggested that Twitter can provide relevant and accurate information related to the outbreak for public authorities and health practitioners.

Desai et al. introduce HurricaneEmo, an emotion dataset of 15,000 English tweets spanning three hurricanes. The authors present a comprehensive study of fine-grained emotions and propose classification tasks to discriminate between coarse-grained emotion groups.

2.3 COVID-19 Emotion Dataset

Since the emergence of the pandemic, numerous studies have been carried out on social media networks to understand COVID-19 and its effects on the larger population. Ils et al. annotated 2.3K German and English tweets for the expression of solidarity and used it to carry out an analysis into the expression of solidarity over time. On the other hand, Saakyan et al. annotated a dataset for detecting general misinformation in the pandemic. Sentiment analysis and emotion detection on social media during COVID-19 have seen tremendous popularity as well due to the ability to provide vital information into the social aspects and the overall dynamics of the population. Besides, Sosea et al. annotate CovidEmo, a dataset of fine-grained emotions and employ a comprehensive analysis into cross-domain and temporal generalization of large pretrained language models. Furthermore, Gupta et al. collected over 198 million Twitter posts from more than 25 million unique users using four keywords: "corona", "wuhan", "nCov" and "covid". The authors label each tweet with seventeen semantic attributes by using topic modeling techniques and pre-trained machine learning-based emotion analytic algorithms.

3. MATERIAL&METHODS

3.1 Initial Explorations

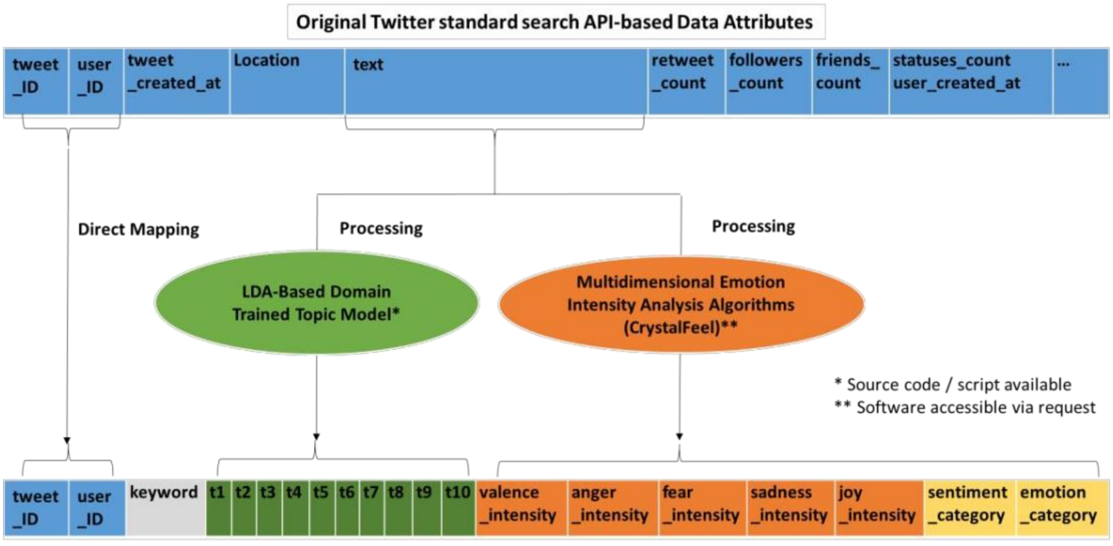
Our initial goal of this project is to create and implement a model that can realize fine-grained sentiment analysis given a text content which may be from social media or news comments. For the fine-grained emotion label, we choose Plutchik’s emotion wheel which suggests 8 primary bipolar emotions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. An important reason of choosing Plutchik’s model instead of Ekman’s six basic emotions is that Ekman’s model was based on some facial expressions, which has large variance with the text generated by natural language.

Ekman’s six basic emotions	Plutchik’s emotion wheel
Anger	Anger
Disgust	Disgust
Fear	Fear
Happiness	Joy
Sadness	Sadness
Surprise	Surprise
	Anticipation
	Trust

Emotions in natural language are complex and each text may have more than one types of emotions. Unlike images or voices, texts may not portray peculiar cues to emotions. Also, the hurdle of detecting emotions from short texts, emojis, and grammatical errors could be back-breaking coupled with the continuous evolution of new words as a result of language dynamics. Therefore, we try to incorporate more complex emotions into our analysis and hence choose Plutchik’s emotion wheel as the emotive model.

Given a specific model, we have to find the dataset which has labels based on this model. The dataset also needs to meet our other constraints, such as the topic or keyword relating to COVID-19. We first search for some COVID-19 emotion datasets, finding that most datasets are labeled by different emotion models that have different sets of labels.

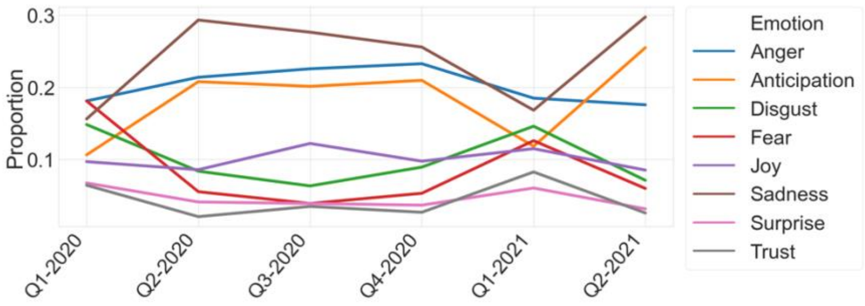
For example, Gupta et al. create a COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes. From 28 January 2020 to 1 September 2021, we collected over 198 million Twitter posts from more than 25 million unique users using four keywords: "corona", "wuhan", "nCov" and "covid". The authors labeled each tweet with seventeen semantic attributes, including a) ten binary attributes indicating the tweet's relevance or irrelevance to the top ten detected topics, b) five quantitative emotion attributes indicating the degree of intensity of the valence or sentiment (from 0: very negative to 1: very positive), and the degree of intensity of fear, anger, happiness and sadness emotions (from 0: not at all to 1: extremely intense), and c) two qualitative attributes indicating the sentiment category (very negative, negative, neutral or mixed, positive, very positive) and the dominant emotion category (fear, anger, happiness, sadness, no specific emotion) the tweet is mainly expressing.



The Present COVID-19 Twitter Dataset with Topics, Sentiments and Emotions Attributes

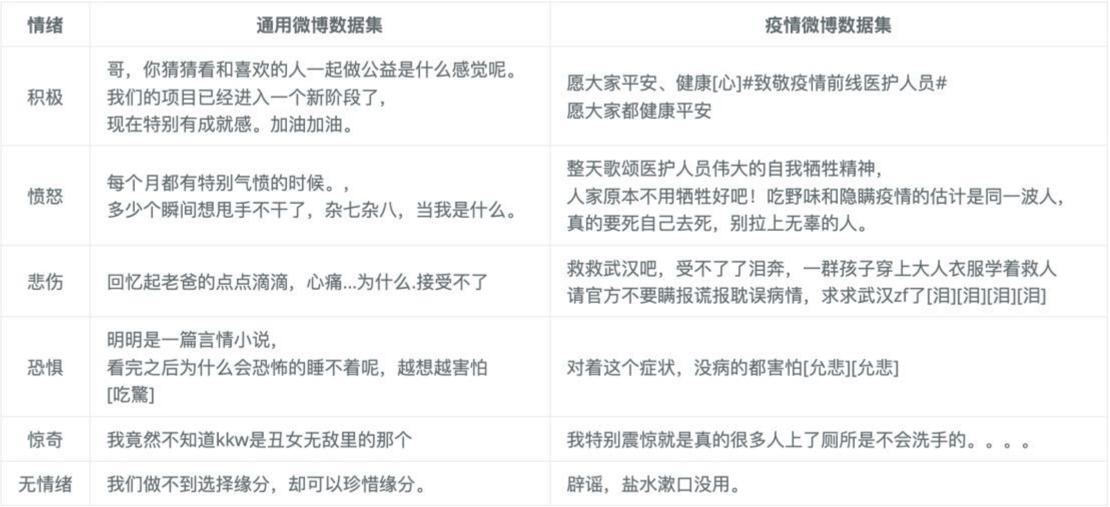
After meticulous searching, we find a twitter dataset under COVID-19 topic and has emotion labels correspond to Plutchik's model. The dataset, created by Sosea et al., is sampled from 129,820 English tweets which is from Chen et al. (2020)'s ongoing collection of tweets related to the COVID-19 pandemic. The authors randomly sample 5, 500 tweets from this data and use Amazon Mechanical Turk to crowdsource Plutchik-8 emotions: anger, anticipation, joy, trust, fear, surprise, sadness, disgust.

Emotion	**Content words/Hashtags**
disgust	**Content words:** disgusting, fucking, million, trump, dead, shit, president, america, china, done **Hashtags:** #hongkong, #gop, #factsmatter, #ccp, #china, #wuhan, #covid19
anger	**Content words:** fuck, evil, bullshit, stupid, idiot, damn, obama, church, lying **Hashtags:** #marr, #covidiots, #trumpvirus, #torycorruption, #skynews, #qanon, #nh, #jacksonville, #gop, #factsmatter
fear	**Content words:** scared, exam, dangerous, infected, confirmed, worse, sir, wuhan, risk, rate **Hashtags:** #stopcovidlies, #jeeneet, #antistudentmodigovt, #health, #wuhan, #china, #stayhome, #covid19
sadness	**Content words:** sad, cry, died, suffering, toll, record, sorry, feel, tested, facing **Hashtags:** #notmychild, #quarantine, #rip, #pregnant, #italy, #healthcare, #freepalestine, #askktr, #wuhan, #vaccine
anticipation	**Content words:** effort, christmas, available, join, start, future, vaccination, vaccinated, coming, open **Hashtags:** #stayhomestaysafe, #pregnant, #postponeinicet, #nyc, #launchzone, #fred2020, #cow, #whatshappeninginmyanmar, #ethereum, #bcpoli
trust	**Content words:** working, support, safe, help, say, being, world, vaccine, good, more **Hashtags:** #stayhome, #staysafe, #covid19, #lockdown, #china
joy	**Content words:** grateful, beautiful, thanks, happy, love, great, little, morning, good **Hashtags:** #taiwan, #innovation, #breaking, #staysafe, #stayathome, #stayhome, #wearamask, #lockdown, #covid19
surprise	**Content words:** believe, year, lockdown, new, china, virus, day, america, covid19, get **Hashtags:** #china, #covid19



However, after downloading the dataset, we find the data to be extremely dirty. The labels of the text were confused because most of them were incorrect based on our manually checking and understanding. For example, text content “Trending on PubMed: Personal knowledge on novel coronavirus pneumonia.” was labeled with “fear”; “Egypt confirms coronavirus case, the first in Africa” was labeled with “surprise”; “Wishing u all a Happy Easter ... Praying for everyone suffering from the corona virus and for everyone trying to help fight this virus... We will overcome this together as one nation united together....” was labeled with “anger”. Obviously, such data cannot be used to train or evaluate.

Those mentioned above are all datasets with English text. For the Chinese part, we only find a COVID-19 emotion dataset from Chinese social media – weibo. It is created by HIT- SCIR and labeled with 6 emotion types: positive, angry, sad, fearful, surprised, and neutral.



情绪	通用微博数据集	疫情微博数据集
积极	哥，你猜猜看和喜欢的人一起做公益是什么感觉呢。 我们的项目已经进入一个新阶段了， 现在特别有成就感。加油加油。	愿大家平安、健康[心]#致敬疫情前线医护人员# 愿大家都健康平安
愤怒	每个月都有特别气愤的时候。， 多少个瞬间想甩手不干了，杂七杂八，当我是什么。	整天歌颂医护人员伟大的自我牺牲精神， 人家原本不用牺牲好吧！吃野味和隐瞒疫情的估计是同一波人， 真的要死自己去死，别拉上无辜的人。
悲伤	回忆起老爸的点点滴滴，心痛...为什么.接受不了	救救武汉吧，受不了了泪奔，一群孩子穿上大人衣服学着救人 请官方不要瞒报谎报耽误病情，求求武汉zf了[泪][泪][泪][泪]
恐惧	明明是一篇言情小说， 看完之后为什么会恐怖的睡不着呢，越想越害怕 [吃惊]	对着这个症状，没病的都害怕[允悲][允悲]
惊奇	我竟然不知道kkw是丑女无敌里的那个	我特别震惊就是真的很多人上了厕所是不会洗手的。。。。
无情绪	我们做不到选择缘分，却可以珍惜缘分。	辟谣，盐水漱口没用。

The dataset is valuable for training the Chinese emotion classifier. However, if we use this dataset to do the fine-grained classification task, we have to find or create an English dataset which has the same label set. So far, we couldn’t find such an appropriate dataset in English. Moreover, this Chinese dataset has no information about the timestamp of each weibo, which means we cannot do the preliminary evaluation of the data and decide whether it is worth continuous exploring. Hence, one of the feasible ways is to collect data from social media using API and label the data manually.

3.2 Data collection

In order to get access to Twitter API, we create a Twitter developer account. In order to get access to Weibo API, we create a Weibo developer account. By using API, we can set our search filter with different constraints to get data. The topic and keywords in English and Chinese are “Wuhan” and “武汉” respectively. It is where COVID-19 first attracts world attention. The timestamp is set to be between Feb 1, 2020 and Feb 29, 2020. The region of the twitter is set to be the U.S. while weibo has no region constraints since almost every user of weibo is Chinese.

We manually label the data with 6 emotion labels: positive, angry, sad, fearful, surprised, and neutral. An N*6 matrix is created that N is the number of sample and each row represents the emotion vector with dimension 6. Each dimension in a row represents one type of emotion (positive, sad, angry, fearful, surprised, and neutral). The value of each dimension is set to 1 if the emotion of the corresponding dimension is the dominant emotion of the text. Otherwise the value is set to 0. Sample data are shown below.

Text	**Emotion Vector**
Officials in Hong Kong are using wristbands to track families that are under Wuhan coronavirus quarantine.	0,0,0,0,0,1 (neutral)
Wuhan new hospital or prison? Iron bar on window, lock on outside, no do...	0,0,0,1,0,0 (fearful)
What a beautiful thread of stories from Wuhan. I love how stories like these make a place far away seem so close.	1,0,0,0,0,0 (positive)
I'm not sure but there's a putrid smoke above Wuhan. #KAG2020	0,0,0,1,0,0 (fearful)

Text	**Emotion Vector**
心系武汉，共战疫情，万众一心！#武汉加油# ，中国加油！	1,0,0,0,0,0 (positive)
武汉有难，是人是狗，心知肚明 除了韩红，其他狗屁明星中国国足中国男篮都是垃圾	0,0,1,0,0,0 (angry)
武汉市市委副书记胡立山:病人没有完全收治,我们很痛苦	0,1,0,0,0,0 (sad)
越想越气,恨透了那些吃野味的人了,强烈希望疫情结束后国家能把贩卖野味和吃野味纳入刑法！	0,0,1,0,0,0 (angry)

3.3 Data Analysis

To ensure our collected data worth to be continuously explored, we do preliminary evaluation on the data. Differences in primary emotion of the text between cultures are expected. Since the emotion vectors are binary, we choose PCA as the method to detect the primary emotion of the texts.

Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data. The principal components are eigenvectors of the data's covariance matrix. Thus, the principal components are often computed by eigen-decomposition of the data covariance matrix or singular value decomposition of the data matrix.

Mathematically, the transformation is defined by a set of size l of p-dimensional vectors of weights or coefficients $\mathbf{w}_{(k)} = (w_1, \dots, w_p)_{(k)}$ that map each row vector $\mathbf{x}_{(i)}$ of X to a new vector of principal component scores $\mathbf{t}_{(i)} = (t_1, \dots, t_l)_{(i)}$, given by

$$t_{k_{(i)}} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)} \quad \text{for } i = 1, \dots, n \quad k = 1, \dots, l$$

in such a way that the individual variables $t_1, \dots, t_l$ of t considered over the data set successively inherit the maximum possible variance from X, with each coefficient vector w constrained to be a unit vector (where l is usually selected to be less than p to reduce dimensionality).

In order to maximize variance, the first weight vector $\mathbf{w}_{(1)}$ thus has to satisfy

$$\mathbf{w}_{(1)} = \arg \underset{\|\mathbf{w}\|=1}{max} \{\sum_i (t_1)^2_{(i)}\} = \arg \underset{\|\mathbf{w}\|=1}{max} \{\sum_i (\mathbf{x}_{(i)} \cdot \mathbf{w})^2\}$$

Equivalently, writing this in matrix form gives

$$\mathbf{w}_{(1)} = \arg \underset{\|\mathbf{w}\|=1}{max} \{\| \mathbf{Xw} \|^2\} = \arg \underset{\|\mathbf{w}\|=1}{max} \{\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Xw}\}$$

Since $\mathbf{w}_{(1)}$ has been defined to be a unit vector, it equivalently also satisfies

$$\mathbf{w}_{(1)} = \arg max\{\frac{\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Xw}}{\mathbf{w}^{\mathrm{T}}\mathbf{w}}\}$$

The quantity to be maximized can be recognized as a Rayleigh quotient. A standard result for a positive semidefinite matrix such as $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is that the quotient's maximum possible value is the largest eigenvalue of the matrix, which occurs when $\mathbf{w}$ is the corresponding eigenvector.

With $\mathbf{w}_{(1)}$ found, the first principal component of a data vector $\mathbf{x}_{(i)}$ can then be given as a score $\mathbf{t}_{1(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}$ in the transformed co-ordinates, or as the corresponding vector in the original variables, $\{\mathbf{x}_{(i)} \cdot \mathbf{w}_{(1)}\}\mathbf{w}_{(1)}$.

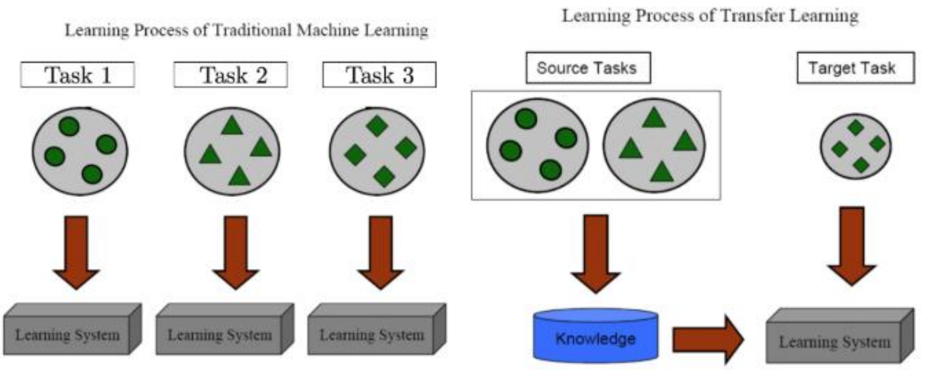
PCA provide us with a method to analyze the principal component (i.e. primary emotion) of the emotion vectors. We only need to split a subset of the data to do the PCA and check whether the difference between cultures is evident.

3.4 Classification Model

3.4.1 Transfer Learning

In these recent times, we have become very good at predicting a very accurate outcome with very good training models. But considering most of the machine learning tasks are domain specific, the trained models usually fail to generalize the conditions that it has never seen before. The real world is not like the trained data set, it contains lot of messy data and the model will make an ill prediction in such condition. The ability to transfer the knowledge of a pre-trained model into a new condition is generally referred to as transfer learning.

Transfer learning is the application gained of one context to another context. So, applying the knowledge from one model could help reduce training time and deep learning issues through taking existing parameters to solve “small” data problems.



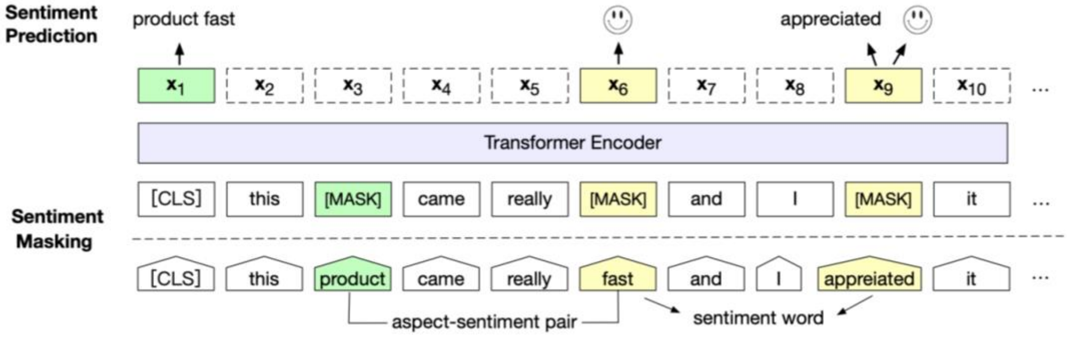
(a) Traditional Machine Learning (b) Transfer Learning

In this study, we use pre-trained model to generate sentence representations, which include context information and emotion information. Then we do fine-tuning in our own dataset to classify fine-grained sentiment.

3.4.2 Pre-trained Model

Pre-training methods have shown their powerfulness in learning general semantic representations and have remarkably improved most natural language processing (NLP) tasks like sentiment analysis. These methods build unsupervised objectives at word-level, such as masking strategy, next-word prediction or permutation. Such word-prediction-based objectives have shown great abilities to capture dependency between words and syntactic structures.

SKEP (Sentiment Knowledge Enhanced Pre-training) is a pre-trained model proposed by Baidu Inc., where sentiment knowledge about words, polarity, and aspect-sentiment pairs are included to guide the process of pre-training. It integrates different types of sentiment knowledge together and provides a unified sentiment representation for various sentiment analysis tasks. This is quite different from traditional sentiment analysis approaches, where different types of sentiment knowledge are often studied separately for specific sentiment tasks.


Sentiment Prediction
product fast
appreciated
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10} ...
Transformer Encoder
[CLS] this [MASK] came really [MASK] and I [MASK] it ...
Sentiment Masking
[CLS] this product came really fast and I appreiated it ...
aspect-sentiment pair
sentiment word

SKEP contains two parts: (1) Sentiment masking recognizes the sentiment information of an input sequence based on automatically-mined sentiment knowledge, and produces a corrupted version by removing these information. (2) Sentiment pre-training objectives require the transformer to recover the removed information from the corrupted version. The three prediction objectives on top are jointly optimized: Sentiment Word (SW) prediction (on $\mathbf{x}_9$), Word Polarity (SP) prediction (on $\mathbf{x}_6$ and $\mathbf{x}_9$), Aspect-Sentiment pairs (AP) prediction (on $\mathbf{x}_1$). Here, the smiley denotes positive polarity. Notably, on $\mathbf{x}_6$, only SP is calculated without SW, as its original word has been predicted in the pair prediction on $\mathbf{x}_1$.

3.4.3 Fine-tuning

Based on the sentiment representation of the text generated by SKEP, the continuous fine-tuning sentiment analysis would be easier. On top of the pre-trained transformer encoder, an output layer is added to perform task-specific prediction. The neural network is then fine-tuned on task-specific labeled data.

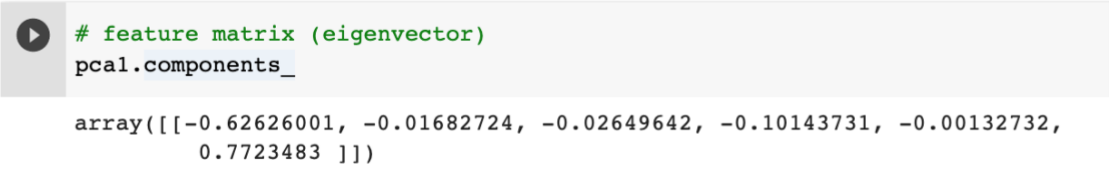
To classify the fine-grained sentiment – (positive, sad, angry, fearful, surprised, neutral), the final state vector of classification token [CLS] is used as the overall representation of an input sentence and is fed into the transformer encoder. On top of the transformer encoder, a classification layer is added to calculate the sentiment probability based on the overall representation.

4. RESULTS

The evaluation of the data is conducted on a subset of 100 samples each language. The region of the English tweets is the U.S. The region of the Chinese weibo is China. Both tweets and weibos have the keyword of same meaning “Wuhan/武汉” respectively. The timestamp of both social media posts is Feb. 5, 2020, which is the same. The labels of the posts are in a set {‘positive’, ‘sad’, ‘angry’, ‘fearful’, ‘surprised’, ‘neutral’}. We use PCA to analyze 3 sets: the concatenated set of tweets and weibos, tweets and weibos respectively, in order to find any similarity or difference in emotion.

4.1 Tweets + Weibos

The emotion representations (vectors) of both tweets and weibos are concatenated on axis 0. The shape of the matrix then turns into 200*6. The second dimension of 6 represents 6 emotion [‘positive’, ‘sad’, ‘angry’, ‘fearful’, ‘surprised’, ‘neutral’] respectively. We use $sklearn.decomposition.PCA$ to build the model and set $n\_components = 1$. After $fit\_transform$ we can check the components (eigenvector) of the PCA model by printing attributes $model.components\_$ to get primary sentiment of the whole set. The result is shown below.



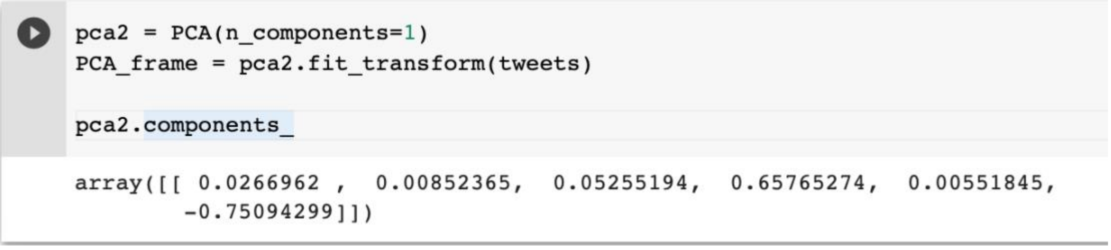
```
# feature matrix (eigenvector)
pca1.components_
```

```
array([[-0.62626001, -0.01682724, -0.02649642, -0.10143731, -0.00132732,
         0.7723483 ]])
```

From the results, we discover that the order of primary emotion is as follow: Neutral > Positive > Fearful > Angry > Sad > Surprised. It’s the order of the primary emotion of the tweets and weibos together.

4.2 Tweets

Then we transform tweets separately. The result is shown below.



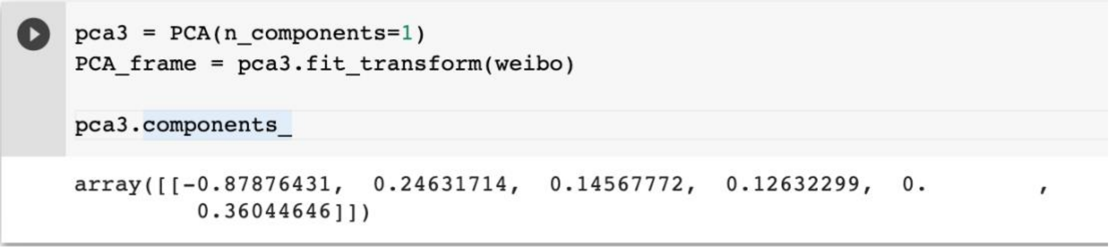
```
pca2 = PCA(n_components=1)
PCA_frame = pca2.fit_transform(tweets)

pca2.components_
```

```
array([[ 0.0266962 ,  0.00852365,  0.05255194,  0.65765274,  0.00551845,
        -0.75094299]])
```

From the results, we discover that the order of primary emotion is as follow: Neutral > Fearful > Angry > Positive > Sad > Surprised. It's the order of the primary emotion of the tweets.

4.3 Weibos

Then we transform weibos separately. The result is shown below.

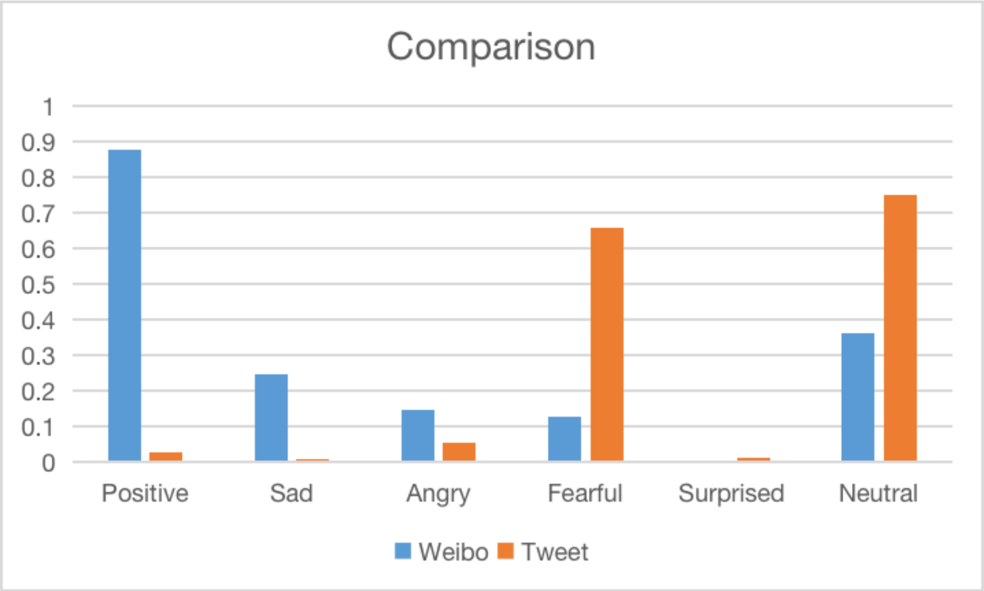


```
pca3 = PCA(n_components=1)
PCA_frame = pca3.fit_transform(weibo)

pca3.components_
```

```
array([[-0.87876431,  0.24631714,  0.14567772,  0.12632299,  0.        ,
         0.36044646]])
```

From the results, we discover that the order of primary emotion is as follow: Positive > Neutral > Sad > Angry > Fearful > Surprised. It's the order of the primary emotion of the weibos.

4.4 Comparison



5. Conclusion

This study aimed to find the differences between sentiments and emotions of the people from the U.S. and China amidst coronavirus (COVID-19) outbreak from their social media posts. Emotional tweet dataset and emotional weibo dataset are used for evaluating fine-grained sentiment from users' posts. Initial tweets and weibos right after the pandemic outbreak were extracted by tracking the trending keyword during February 2020. The subset for preliminary evaluation was extracted by a specific timestamp - Feb. 5, 2020.

Our findings showed a high percentage of social media posts from both the U.S. and China had "Neutral" label that no evident emotion was revealed. A possible reason for this is that both Twitter and Weibo are one of the platforms for media to post news and neutral comments. Besides, people from the U.S. seemed to express their feeling more neutrally.

Whereas, despite the similarity in 'Neutral' sentiment, the tweets and weibos posted following the corona outbreak between two countries showed quite different trend of emotion. "Positive" emotion tended to be the most dominant emotion in weibos while it less frequently appeared in tweets. This may be due to Chinese people tended to use words "加油", "好起来", "祝福", "温暖", "正能量", which expressed their encouragement and blessing to the people who suffered in the pandemic. Moreover, Chinese seemed very empathetic to the touching event related to COVID-19 and thus posted more content that they are moved. This also may be the reason why the "Sad" emotion was relatively high in weibos compared to other emotions except "Positive" and "Neutral". People in the U.S. seldom revealed "Sad" feeling in social media. It can be partly explained by the fact that the pandemic was not so serious, or say, was not taken seriously in the U.S. (there were lots of COVID-19 cases already but not yet detected). Therefore, most people were not feeling sad about the loss caused by the pandemic around them.

As a matter of fact, a dominant emotion of tweets at that time was "Fearful". Because of the spread, people in the U.S. were afraid of the infection going more and more quickly that they themselves would be in danger soon. On the contrary, Chinese people had less "Fearful" emotion since the pandemic had already existed and the government had already taken a series of action to stop the spread.

The "Angry" emotion was expressed partly because of the dissatisfaction to the people who might cause the COVID-19 to happen and partly because the action taken by the government was not so satisfying. More "Angry" emotion in China because the pandemic first came into our view from Wuhan, and Chinese government took more actions than other countries at that time.

People in both the U.S. and China hardly had any "Surprised" emotion according to the evaluation of the subset. We infer this was because the COVID-19 had caught people's attention for about one month at that time. Hence people got used to the news about the pandemic and media hadn't used shocking words already. On the other hand, governments were trying not to spread the panic over the society.

In conclusion, people in the U.S. and China had many differences in expressing emotion on social media, which met our expectations. Our work on this topic is worth continuous exploring.

6. FUTURE WORK

6.1 Labeling More Data

From the results and conclusion above, we discover that the method we used to collect data was valid and the topic is worth further exploring. In order to train a fine-grained emotion classifier based on pre-trained model, more data is needed. Hence labeling more data would be the first work in our next step. Apart from labeling data manually, there are some unsupervised and semi-supervised methods that can label the data automatically. We should find an efficient way to expand our dataset.

6.2 Implementing Fine-Grained Emotion Classifier

For the pre-trained model part, we choose SKEP as the basic model to generate emotion representation of the text. Changing the structure of the model is required in order to do the fine-tuning. On top of the pre-trained transformer encoder, adding an output layer is needed to perform task-specific prediction. The neural network is then fine-tuned on task-specific labeled data. As long as we have enough data, we would be able to train the model to be a fine-grained emotion classifier.

6.3 Further Analysis of Data in Time Series

After training the classifier, more and more social media posts or news comments can be analyzed in sentiment. Not only the posts in Twitter and Weibo can be analyzed, but also the related news and video comments, which provide possibility for determining which textual tags are preferred by different affinity groups for news and related videos. Furthermore, the difference between cultures in sentiment may vary as time goes on. The change of the cross-culture sentiment may also be interesting to explore and study.

REFERENCE

[1] Nisbett, R. (2004). The geography of thought: How Asians and Westerners think differently... and why. Simon and Schuster.

[2] Ekman, P., & Keltner, D. (1997). Universal facial expressions of emotion. Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture, 27, 46.

[3] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of personality and social psychology, 17(2), 124.

[4] Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American scientist, 89(4), 344-350.

[5] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In A practical guide to sentiment analysis (pp. 1-10). Springer, Cham.

[6] Liang, H., Fung, I. C. H., Tse, Z. T. H., Yin, J., Chan, C. H., Pechta, L. E., ... & Fu, K. W. (2019). How did Ebola information spread on twitter: broadcasting or viral spreading? BMC public health, 19(1), 1-11.

[7] Prabhakar Kaila, D., & Prasad, D. A. (2020). Informational flow on Twitter–Corona virus outbreak–topic modelling approach. International Journal of Advanced Research in Engineering and Technology (IJARET), 11(3).

[8] Fu, K. W., Liang, H., Saroha, N., Tse, Z. T. H., Ip, P., & Fung, I. C. H. (2016). How people react to Zika virus outbreaks on Twitter? A computational content analysis. American journal of infection control, 44(12), 1700-1702.

[9] Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS one, 5(11), e14118.

[10] Hasan, M., Rundensteiner, E., & Agu, E. (2014). Emotex: Detecting emotions in twitter messages.

[11] Fung, I. C. H., Tse, Z. T. H., Cheung, C. N., Miu, A. S., & Fu, K. W. (2014). Ebola and the social media.

[12] Desai, S., Caragea, C., & Li, J. J. (2020). Detecting perceived emotions in hurricane disasters. arXiv preprint arXiv:2004.14299.

[13] Ils, A., Liu, D., Grunow, D., & Eger, S. (2021). Changes in European Solidarity Before and During COVID-19: Evidence from a Large Crowd-and Expert-Annotated Twitter Dataset. arXiv preprint arXiv:2108.01042.

[14] Saakyan, A., Chakrabarty, T., & Muresan, S. (2021). COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. arXiv preprint arXiv:2106.03794.

[15] Sosea, T., Pham, C., Tekle, A., Caragea, C., & Li, J. J. (2021). Emotion analysis and detection during COVID-19. arXiv preprint arXiv:2107.11020.

[16] Gupta, R. K., Vishwanath, A., & Yang, Y. (2020). Global Reactions to COVID-19 on Twitter: A Labelled Dataset with Latent Topic, Sentiment and Emotion Attributes. arXiv preprint arXiv:2007.06954.

[17] Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., ... & Wu, F. (2020). SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. arXiv preprint arXiv:2005.05635.

[18] Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. (2020). Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. IEEE Access, 8, 181074-181090.

[19] Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2(7), e12189.

[20] Müller, M., Salathé, M., & Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503.

[21] Li, J., Xu, Q., Cuomo, R., Purushothaman, V., & Mackey, T. (2020). Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: retrospective observational infoveillance study. JMIR Public Health and Surveillance, 6(2), e18700.

[22] Wang, T., Lu, K., Chow, K. P., & Zhu, Q. (2020). COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. Ieee Access, 8, 138162-138169.

[23] Yu, X., Zhong, C., Li, D., & Xu, W. (2020, November). Sentiment analysis for news and social media in COVID-19. In Proceedings of the 6th ACM SIGSPATIAL International Workshop on Emergency Management using GIS (pp. 1-4).