National Tsing Hua University

--2021 ISA 5810 Data Mining

DM 2021 ISA5810

Lab 2 Homework

Emotion recognition on twitter

Name: 陳翼弘
Student ID: 103033617
GitHub ID: YiHungChen
Kaggle name 22215314

Chapter 1. Data preprocessing

1.1 Removing extra characters

From the original text data in the tweets, it is easy to find that when people are tweets they tend to use more than just regular words or phrases. Novel characters or symbols are the favored choice when the sentence needs to be short and powerful. However, some of the symbols are meaningless such as '<LH>' or '___'. These characters are removed during the data cleaning process.

1.2 Lemmatization

After removing the extra characters, the next step is to normalize the tense, plural noun, and contraction from the apostrophe. The normalized tense and the plural noun are being achieved by the library of WordNetLemmatizer from nltk. The contraction from the apostrophe is being replaced with the original forms of nouns and verbs. The rules are shown in Table 1. Table 2 shows the example of the articles before and after the lemmatization process.

Table 1 rule of contraction

's	→	is
've	→	have
n't	→	not
're	→	are
'm	→	am
'll	→	will

Table 2 Articles before and after the lemmatization

ID	Text	Lemmatization
0x376b20	People who post "add me on #Snapchat" must be dehydrated. Cuz man.... that's <LH>	People who post "add me on Snapchat" must be dehydrate . Cuz man .... that be
0x359db9	The #SSM debate; <LH> (a manufactured fantasy used to distract the ignorant masses from their mundane lives) V #gender #diversity (a m......	The SSM debate; (a manufacture fantasy use to distract the ignorant mass from their mundane life) V gender diversity (a m ......

Chapter 2. Feature extraction

The idea of this project is to describe emotion as a color. In this competition, there are eight different emotions as [anger, joy, anticipation, disgust, fear, sadness, surprise, trust]. By the concept of color, every word is assumed to compose of eight different emotions (color). A simple example is considered only with three different emotions. In this case, three different emotions are regarded as the color of R, G, and B. Then the color space RGB can be converted to the HSV space, which is using three other components to describe the color. The benefit of the HSV space is that it can easily tell the difference from one color to another. Back to this project, the idea of the HSV is also believed can be taken as the feature of emotion classification. However, the H in HSV space stands for the components of Hue, which is a value of the angle. This value only works with three-dimensional features. So, this value is useless for the eight-category cases. The value S stands for saturation. This value is used to tell how the color is outstanding from the others. Or in other words, how colorful it is. And the last value is V, which stands for the value or intensity. This value is to describe how strong the color is.

2.1 The intensity of emotion extraction

The intensity of words in different emotions is calculated by how many times those words appear in the emotion category. For example, if the words [love] appears in the category {Joy} for 743 times. The intensity of the emotion {Joy} in word [love] then is 743. Then calculate the other emotions and processes with normalization. Table 3 shows the motion intensity from the words [love] and [be]. Table 4 shows the motion intensity with the normalization from the words[love] and [be]. Also, Figure 1shows the emotion ratio figure of words [love] and [be].

Table 3 intensity of the words [love] and [be]

	Anger	Joy	Anticipation	Disgust	Fear	Sadness	Surprise	Trust
love	76	743	246	82	150	130	142	245
be	2482	3245	3272	3125	2462	2682	3091	2736

Table 4 intensity of the words after the normalization

	Anger	Joy	Anticipation	Disgust	Fear	Sadness	Surprise	Trust
love	0	1	0.25	0.00	0.11	0.08	0.09	0.25
be	0.024	0.96	1	0.81	0	0.27	0.776	0.338

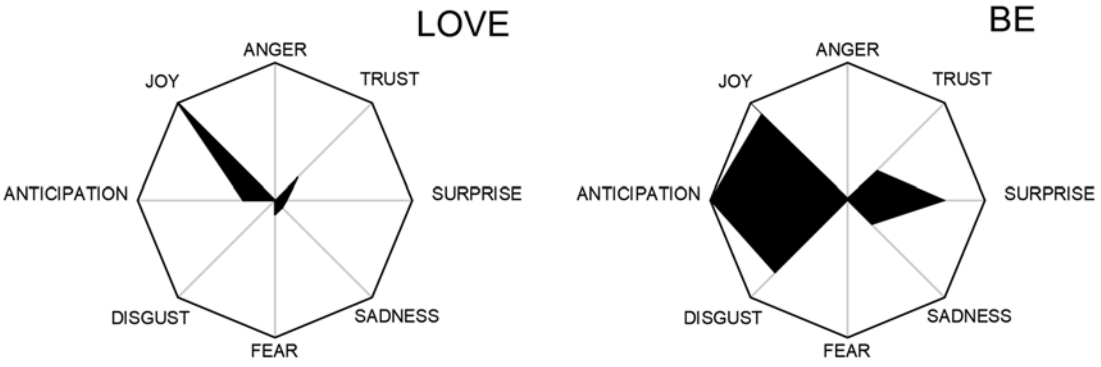

Figure 1 emotion ratio figure of [love] and [be]

2.2 Calculation of emotion saturation

From the previous section, the intensity of the emotion can be revealed. However, higher emotion intensity does not necessarily connect to the better representation of the emotion. There might exist multiple emotions with strong intensity. Take the word [BE] in FIGURE XXX as an example, the word [BE] is frequently being used in the category of [JOY], [ANTICIPATION], and [DISGUST]. Considering articles with [BE] as any of the categories is risky. To solve this problem, an evaluation value of the importance of the words is required. In this project, the idea of saturation from HSV is being applied. Eq. 1 shows the original formula of the calculation of the saturation. C_{max} is the maximum value of the R, G, B; C_{min} is the minimum value of the R, G, B.

$$\frac{C_{max} - C_{min}}{C_{max}} \qquad \text{Eq. 1}$$

By considering the saturation from another perception, the calculation of the saturation is considered as the ratio between the maximum value and the difference from the third large value. The larger the saturation means the maximum color is more different or more outstanding from the third color (last color in case of RGB). This also means that the saturation value allows the object to have two colors with strong intensity. Now if the color is replaced by the emotion and the dimension is increased from 3 to 8. The calculation of the saturation is shown in Eq. 2. C_{nmax} is the n^{th} large emotion intensity of the word.

$$\frac{C_{max} - C_{nmin}}{C_{max}} \qquad \text{Eq. 2}$$

By adjusting the number of n, it is possible to increase or decrease the tolerance of how many emotions are allowed in a single word. Table 5 shows the result of calculating the saturation. And Table 6 and Figure 2 show the intensity and the emotion ratio figure after multiplying the intensity with the saturation.

Table 5 saturation of words [love] and [be]

	Anger	**Joy**	**Anti.**	**Disgust**	**Fear**	**Sadness**	**Surprise**	**Trust**	**Sat.**
love	0	1	0.25	0.00	0.11	0.08	0.09	0.25	0.668
be	0.024	0.96	1	0.81	0	0.27	0.776	0.338	0.008

Table 6 intensity after multiplied by the saturation

	Anger	**Joy**	**Anti.**	**Disgust**	**Fear**	**Sadness**	**Surprise**	**Trust**	**Sat.**
love	0.000	0.669	0.170	0.006	0.074	0.054	0.066	0.169	0.668
be	0.0002	0.007	0.008	0.006	0	0.002	0.006	0.002	0.008

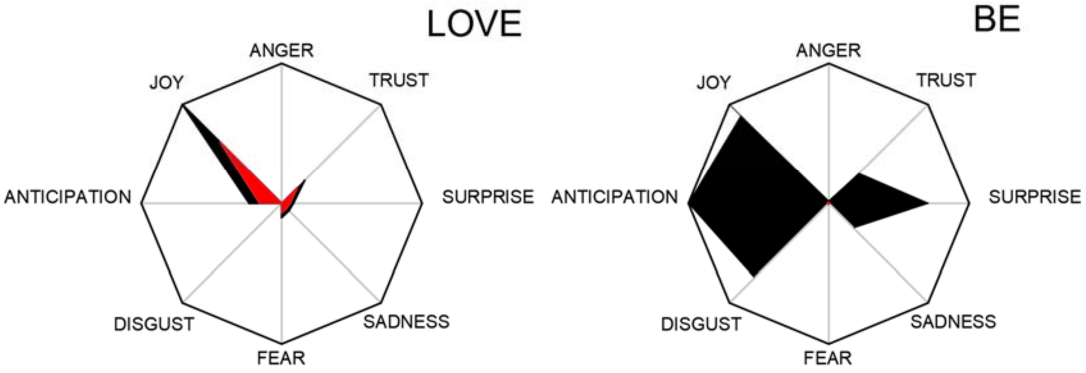

Figure 2 emotion ratio figure after multiplying the saturation. (Black: original value, Red: After multiply saturation)

Chapter 3. Classification:

3.1 Emotion score

Before applying the classification method, a very quick question is that right now the emotion intensity value after multiplying by the saturation is been calculated. Is it possible to calculate the result by simply comparing the value of the word emotion intensity from the article? This method is been carried out by summing up all the words emotion intensity in the article. Table 7 shows the emotion score table for a tweet. The result indicates that the emotion of this tweet should be surprise.

Table 7 emotion score table for tweet

	Anger	**Joy**	**Anti**	**Disgust**	**Fear**	**Sadness**	**Surprise**	**Trust**
add	0.0843	0.1855	0.1012	0.1349	0.0000	0.1349	0.4722	0.1181
be	0.0002	0.0080	0.0083	0.0068	0.0000	0.0022	0.0064	0.0028
cuz	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
dehydrate	0.2500	0.5000	0.2500	0.2500	0.0000	0.2500	0.0000	0.2500
man	0.0000	0.0601	0.0769	0.0529	0.0433	0.0216	0.0240	0.0120
me	0.0774	0.1353	0.0780	0.0000	0.0000	0.0019	0.0887	0.0673
must	0.0000	0.0490	0.2941	0.1103	0.0980	0.1103	0.0245	0.0858
on	0.0424	0.0629	0.0731	0.1580	0.0762	0.0976	0.1007	0.0000
people	0.0735	0.0418	0.0400	0.0855	0.0000	0.0496	0.0550	0.0353
post	0.0532	0.0399	0.0000	0.0976	0.0044	0.0576	0.0355	0.0798
snapchat	0.0537	0.0671	0.0000	0.0939	0.0000	0.0671	0.8182	0.0537
that	0.0000	0.0076	0.0217	0.0274	0.0105	0.0136	0.0300	0.0000
who	0.0040	0.0113	0.0313	0.0269	0.0000	0.0089	0.0294	0.0221
RESULT	0.6388	1.1685	0.9746	1.0440	0.2324	0.8154	**1.6846**	0.7268

The intensity and the saturation not only can calculate the emotion score table for the tweets but also can be applied as a threshold like the TF-IDF. It is possible to select a limited number of words based on their intensity and saturation. This method supports a new way to evaluate and generate the word list for the BOW method.

After receiving the word list or the emotion score of each tweet, further classification methods can be applied to these features. In this research, the author used Naïve Bayesian, Decision Tree, Logistic regression, and neural network.

3.2 Emotion image

The previous section indicated that there can be 10 – 15 words in a tweet. In total, there are 10 * 8 features in one article. However, during the calculation of the emotion score, the feature is compressed into the dimension of eight. A lot of features are expected lost in the process. On the other hand, when humans are using words to describe the emotion, with different combinations of the words, it might result in the different emotion categories. To preserve that information, the idea with a convolution neural network is been proposed. Instead of summing up all the emotion scores in the tweets, the method of emotion image preserved all the values, and save the table into a 30 * 8 figure. Those pixels without the words are considered zero. Figure 3 shows an emotional image for one of the training data.

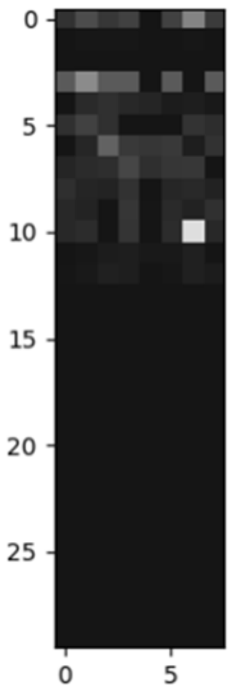

Figure 3 emotion image.

After calculating the emotion image, the image is provided as the input feature for the convolution neural network.

Chapter 4. Result and Discussion

4.1 Experiment result

Table 8 Result of different methods.Table 8 shows the result of emotion classification with different processing methods. The selected result is using the word list filtered with emotional intensity and emotion saturation and the naïve Bayesian as the classification method. The last method is using BIRT to create a state of arts.

Table 8 Result of different methods.

	Data preprocessing	**Feature Extraction**	**Feature processing**	**Classification**	**Result**
1	Lemmatize	Emotion extraction	Emotion score	N/A	0.319
2				Decision Tree	0.294
3				Neural network	0.387
4				Logistic regression	0.401
5			Wordlist & BOW	Decision Tree	0.289
6				Neural network	0.407
7				Logistic regression	0.423
8				Naïve Bayesian	**0.423**
9			Emotion image	CNN	0.418
10		N/A	N/A	BIRT	**0.496**

4.2 Discussion

Although the idea of emotional intensity and saturation is seen as solid, the result tells another story. Two possible reasons are proposed. They are the data unbalance and the data overfitting.

4.2.1 Data unbalance

Since the emotional intensity of each word is calculated by the appearance quantity in every emotion category. The result of the intensity is highly related to the sampling method. From the training dataset, the ratio of tweets in each emotion category is shown in Figure 4.

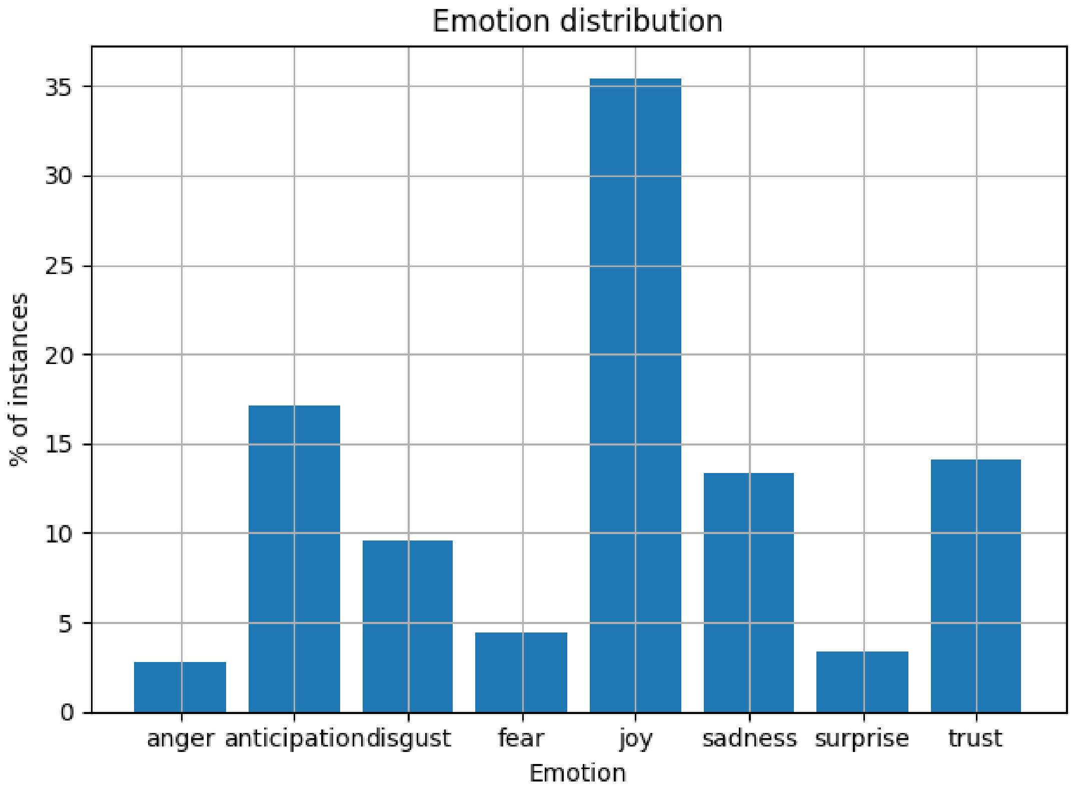

Figure 4 emotion distribution

From figure 4, the category joy is the largest dataset in the training data. It has almost eight times larger than the category of anger. If the sampling process preserves the ratio of the categories, then a higher quantity of articles in joy can be expected. In this case, the emotion score of joy is possible to have higher intensity and result in the wrong categories. To solve this problem, sampling the dataset becomes a critical issue.

4.2.2 Overfitting

Another issue associated with the sampling process is the overfitting problem. The emotion score system is working under the condition of all the provided words are recognized at least once in the training data. To make the largest probability of recognizing as many words as possible. Full training data must be put into the emotion score calculating process at least for one time. Otherwise, some testing articles may have no emotion score. However, according to the previous section, a full training dataset will result in a data unbalanced result. So, the issue becomes the struggle force between the unbalance and overfitting issue.

Chapter 5. Conclusion and Future work

From Chapter 4, it is indicated that the emotion score system is highly related to the words list and the data balance property. However, trying to increase either the quantity of the words lists or the data balance property with only the training dataset will result in the lower performance of the other characteristics. One possible solution may be having a pre-training data set that is large enough to provide complete word lists and also not related to the existing training dataset. In this assumption, the emotional intensity and the saturation can be processed with the pre-trained dataset. And the training dataset is only provided to train the classification model or calculate the emotion score.