Team project #2

The summary of your team project consists of the following parts:

1. The title of your project

Using BERT and Ruled Based Sentiment Analysis to Predict PTT Public Opinion Responding Level (透過BERT與Ruled Based進行多標籤極性分析以預測輿論影響程度)

2. Your team member names

ll0078501 李念紘 Joanne Lee

ll0078506 邱可人 Karen Chiu

ll0078509 施宇軒 Leo Shr

3. The source of text documents, how do you get them.

First, we searched for some famous Taiwanese athletes on PTT and choose five athletes which represent five different sports. We used BeautifulSoup and a for loop to scrape and set the number of pages(in our final project we set five pages for each athlete) on the PTT Gossiping and got the titles of each article. There are twenty articles on each page, which means we got one hundred articles for each athlete. Then, we used Beautifulsoup and a for loop to capture the content, comments, and comment number of each article through the title stored them in the list and then wrote them into a .csv file separately.

4. The sequence of processing the text

1. Data cleaning
   An article of an athlete is a document. We collected a lot of documents and did some data preprocessing on them. First, we added new terms for tokenization and tokenized those words through CKIP, and we also defined new stop words and removed those stop words. We applied NER to tag the part of speech and only kept the adjectives, nouns and proper nouns.
2. BERT
   We adopted BERT as our unsupervised learning method for the training dataset. We used articles about athletes as the input and want to get the result of positive or negative sentiment. But the difficulty we face was the data we have were too little. Thus, we used transfer learning which was the trained model from other professionals to help us finish training our own data and then got the result.
3. Keymoji
   We used Keymoji to help us process Rule-Based optimization of our documents. KeyMoji estimates are based on the eight major paired polar core emotions proposed by American psychologist Robert Plutchik, namely Anger, Anticipation, Disgust, Fear, Joy, Sadness, Sur

prise, and Trust. First, we only use the titles of each article as our input, but the result cam e out with a lot of 0. We guessed that the sentiment score could not be effectively obtained due to the small vocabulary of the title. Therefore, we add content to our input to make our document longer and enrich sentiment scores in our results.

4. Decision Tree
   First, we use the Decision Tree as our training model to determine the relationship betwee n the eight Keymoji sentiment scores and the "negative" or "positive" category.
5. Random Forest
   Bagging Decision Tree are too stable, we applied Random Forest (include Bagged Decisio n Tree & Drop out) as our another prediction model to predict how many comments an arti cle has.
6. Predictive Model Selection
   Which model or skill to apply is critical for accuracy, these are called the tuning skill. To a ccomplish this goal, we built customized functions to execute the techniques. We embrace d K-fold Validation, bagging, and boosting to check the models' RMSE (Root-Mean-Squ are Error) which became the basis for our selection of models.
7.

5. The objectives of the text processing; that is, how would you like to use the outputs for?

First, we want to understand how the Taiwanese think of those famous athletes, and how t hey would describe them (Text Cloud). Go one step further, we want to use the sentiment score th at appear in the titles and the content of the article to predict PTT public opinion responding level, which means how many comments an article has. Also, we try to find out which sentiments affect the number of comments the most. The result can give some suggestions to those who want to attr act people's attention like Youtuber or journalists, etc.

6. The output of the results.

- Decision Tree + Kimoji
  According to our result in Decision Tree, we find out the relationship between eight Key moji sentiment indicators and the "negative" or "positive" category:
  1. Most of the Gossiping version articles have less relationship with sadness and are i n line with common sense.
  2. Joy has the highest average value among the sentiment indicators, which is in line with the characteristics of the Gossiping version.
  3. Joy and Anger represent pride occupying most of the right half of the picture. We speculate that it may be related to the training materials we use, which were article s of well-known athletes, and the villager culture of PTT.
  4. Anger and Disgusting represent the emotion of contempt, occupying most of the ar ticles with Anger emotion.
  5. Because the Gini coefficient is used for segmentation, the trust score is the most ev enly distributed among the sentiment indicators in each segment. Therefore, 4 of th

e 8 layers are classified using Trust.

- Feature Importance
  According to our result in Random Forest, we find out that the most important sentiment f eature to affect the number of comments are **Anticipation**, and **Surprise**, which then is **Fea r**.
- Prediction
  Through our prediction, we can now use sentiment scores to predict the number of comme nts on a post.
-

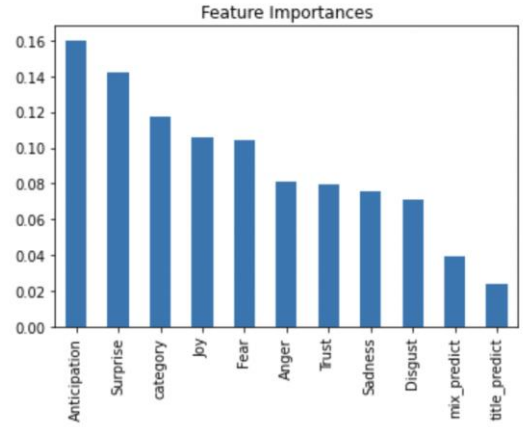
7. **The self-evaluation of the work, including the contributions of the results, the lessons learned, a nd the potential application domains.**

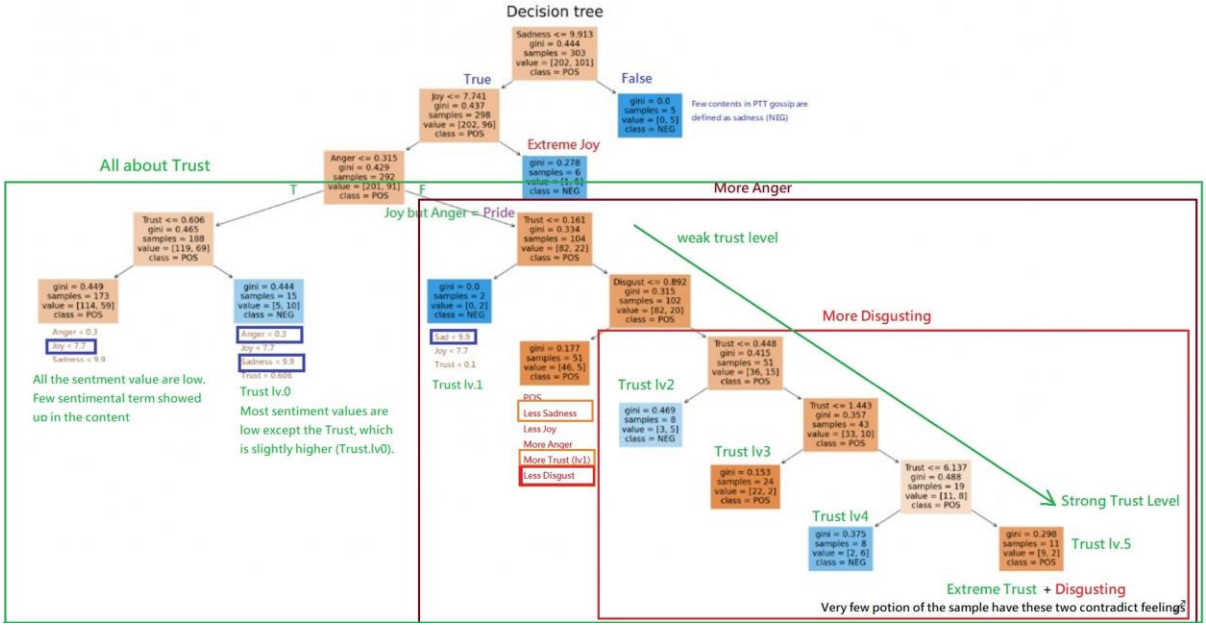
We have used BERT and Keymoji to get the sentiment score of each article, then tried the supervised learning approach to predict the number of comments. First, BERT has some restriction s like a word limit so we have to do more data cleaning in the beginning. Secondly, we used the fr ee version of Keymoji which have a usage limit, so we have to spend more time and cut our docu ment into a smaller dataset as our input. In the future, we should find some way to deal with this p roblem or write code to increase our efficiency.

Also, the problem we met is our data were too little which lead to our prediction accuracy not being high enough. So in the future, we can extend the number of documents, for example, we can grasp 100 pages of articles on the Gossip page in PTT, which is about 2000 articles.

Finally, in the future, we can further enhance our model. For example, we can add mediato rs to the model, or have a discussion about interaction.

8. **Your link to your codes that TA and I can access.**


Feature Importances
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Anticipation
Surprise
category
Joy
Fear
Anger
Trust
Sadness
Disgust
mix_predict
title_predict


Decision tree
Sadness <= 9.913
gini = 0.444
samples = 303
value = [202, 101]
class = POS
True
False
gini = 0.0
samples = 5
value = [0, 5]
class = NEG
Few contents in PTT gossip are defined as sadness (NEG)
Joy <= 7.741
gini = 0.437
samples = 298
value = [202, 96]
class = POS
Extreme Joy
gini = 0.278
samples = 6
value = [1, 5]
class = NEG
All about Trust
Anger <= 0.315
gini = 0.429
samples = 292
value = [201, 91]
class = POS
T
F
Joy but Anger = Pride
More Anger
Trust <= 0.606
gini = 0.465
samples = 188
value = [119, 69]
class = POS
Trust <= 0.161
gini = 0.334
samples = 104
value = [82, 22]
class = POS
weak trust level
gini = 0.449
samples = 173
value = [114, 59]
class = POS
Anger < 0.3
Joy < 7.7
Sadness < 9.9
gini = 0.444
samples = 15
value = [5, 10]
class = NEG
Anger < 0.3
Joy < 7.7
Sadness < 9.9
Trust < 0.606
gini = 0.0
samples = 2
value = [0, 2]
class = NEG
Sad < 9.9
Joy < 7.7
Trust < 0.1
Disgust <= 0.892
gini = 0.315
samples = 102
value = [82, 20]
class = POS
More Disgusting
All the sentment value are low. Few sentimental term showed up in the content
Trust lv.0
Most sentiment values are low except the Trust, which is slightly higher (Trust.lv0).
Trust lv.1
gini = 0.177
samples = 51
value = [46, 5]
class = POS
POS
Less Sadness
Less Joy
More Anger
More Trust (lv1)
Less Disgust
Trust <= 0.448
gini = 0.415
samples = 51
value = [36, 15]
class = POS
Trust lv2
gini = 0.469
samples = 8
value = [3, 5]
class = NEG
Trust <= 1.443
gini = 0.357
samples = 43
value = [33, 10]
class = POS
Trust lv3
gini = 0.153
samples = 24
value = [22, 2]
class = POS
Trust <= 6.137
gini = 0.488
samples = 19
value = [11, 8]
class = POS
Strong Trust Level
Trust lv4
gini = 0.375
samples = 8
value = [2, 6]
class = NEG
gini = 0.298
samples = 11
value = [9, 2]
class = POS
Trust lv.5
Extreme Trust + Disgusting
Very few potion of the sample have these two contradict feelings