# Kaggle Competition Report

113-1 Data Mining Lab 2

江信彥 CHIANG Hsin-Yen (NTNU)

December 8, 2024

## 1 Introduction

In this competition, our task is to predict the emotion of each given tweet id. The dataset is separated into several files: data_identification.csv identifies the membership of the training or testing set for each tweet_id; tweets_DM.json provides the raw data directly from Twitter; and emotion.csv indicates the emotion label for each tweet ID in the training dataset.

We need to submit a file that includes the tweet ID and the predicted emotion of each piece of data in the testing dataset.

## 2 Preprocessing

To train the model, we need to get the text part of the raw data from the tweet in the JSON file. The following figure shows the result of the preliminary processing:

| | _score | _index | _source | _crawldate | _type |
|---|---|---|---|---|---|
| 0 | 391 | hashtag_tweets | {'tweet': {'hashtags': ['Snapchat'], 'tweet_id... | 2015-05-23 11:42:47 | tweets |
| 1 | 433 | hashtag_tweets | {'tweet': {'hashtags': ['freepress', 'TrumpLeg... | 2016-01-28 04:52:09 | tweets |
| 2 | 232 | hashtag_tweets | {'tweet': {'hashtags': ['bibleverse'], 'tweet_... | 2017-12-25 04:39:20 | tweets |
| 3 | 376 | hashtag_tweets | {'tweet': {'hashtags': [], 'tweet_id': '0×1cd5... | 2016-01-24 23:53:05 | tweets |
| 4 | 989 | hashtag_tweets | {'tweet': {'hashtags': [], 'tweet_id': '0×2de2... | 2016-01-08 17:18:59 | tweets |
| ... | ... | ... | ... | ... | ... |
| 1867530 | 827 | hashtag_tweets | {'tweet': {'hashtags': ['mixedfeeling', 'butim... | 2015-05-12 12:51:52 | tweets |
| 1867531 | 368 | hashtag_tweets | {'tweet': {'hashtags': [], 'tweet_id': '0×29d0... | 2017-10-02 17:54:04 | tweets |
| 1867532 | 498 | hashtag_tweets | {'tweet': {'hashtags': [], 'tweet_id': '0×2a6a... | 2016-10-10 11:04:32 | tweets |
| 1867533 | 840 | hashtag_tweets | {'tweet': {'hashtags': [], 'tweet_id': '0×24fa... | 2016-09-02 14:25:06 | tweets |
| 1867534 | 360 | hashtag_tweets | {'tweet': {'hashtags': ['Sundayvibes'], 'tweet... | 2016-11-16 01:40:07 | tweets |

1867535 rows × 5 columns

The columns _index and _type have the same values for all rows. In addition, the _score and _crawldate columns contain the tweet's metadata, which is unrelated to the tweet's content. Next, we further process the data in the _source column:

| | tweet_id | text |
|---|---|---|
| 0 | 0×376b20 | People who post "add me on #Snapchat" must be ... |
| 1 | 0×2d5350 | @brianklaas As we see, Trump is dangerous to #... |
| 2 | 0×28b412 | Confident of your obedience, I write to you, k... |
| 3 | 0×1cd5b0 | Now ISSA is stalking Tasha 😭😭😭 <LH> |
| 4 | 0×2de201 | "Trust is not the same as faith. A friend is s... |
| ... | ... | ... |
| 1867530 | 0×316b80 | When you buy the last 2 tickets remaining for ... |
| 1867531 | 0×29d0cb | I swear all this hard work gone pay off one da... |
| 1867532 | 0×2a6a4f | @Parcel2Go no card left when I wasn't in so I ... |
| 1867533 | 0×24faed | Ah, corporate life, where you can date <LH> us... |
| 1867534 | 0×34be8c | Blessed to be living #Sundayvibes <LH> |

1867535 rows × 2 columns

Now, we obtain a dataframe that includes only the tweet_id and the text.

# 3 Data Cleaning

Note that the text of each tweet includes <LH> tags, emojis, and mention tags. We transform the emoji into text and delete others since they are not related to the emotion of the content.

# 4 Model

## 4.1 Preparation of Training and Testing Datasets

Before training, we need to separate the training dataset and the testing dataset specified in the data_identification.csv file, and merge the emotion label with the corresponding tweet in the training dataset. The following figures show the result:

| | tweet_id | text | identification | emotion |
|---|---|---|---|---|
| 0 | 0×376b20 | People who post "add me on #Snapchat" must be ... | train | anticipation |
| 1 | 0×2d5350 | @brianklaas As we see, Trump is dangerous to #... | train | sadness |
| 2 | 0×1cd5b0 | Now ISSA is stalking Tasha [joy][joy][joy] | train | fear |
| 3 | 0×1d755c | @RISKshow @TheKevinAllison Thx for the BEST TI... | train | joy |
| 4 | 0×2c91a8 | Still waiting on those supplies Liscus. | train | anticipation |
| ... | ... | ... | ... | ... |
| 1455558 | 0×321566 | I'm SO HAPPY!!! #NoWonder the name of this sho... | train | joy |
| 1455559 | 0×38959e | In every circumtance I'd like to be thankful t... | train | joy |
| 1455560 | 0×2cbca6 | there's currently two girls walking around the... | train | joy |
| 1455561 | 0×24faed | Ah, corporate life, where you can date using ... | train | joy |
| 1455562 | 0×34be8c | Blessed to be living #Sundayvibes | train | joy |

1449816 rows × 4 columns

| | tweet_id | text | identification |
|---|---|---|---|
| 2 | 0×28b412 | Confident of your obedience, I write to you, k... | test |
| 4 | 0×2de201 | "Trust is not the same as faith. A friend is s... | test |
| 9 | 0×218443 | When do you have enough ? When are you satisfi... | test |
| 30 | 0×2939d5 | God woke you up, now chase the day #GodsPlan #... | test |
| 33 | 0×26289a | In these tough times, who do YOU turn to as yo... | test |
| ... | ... | ... | ... |
| 1867525 | 0×2913b4 | "For this is the message that ye heard from th... | test |
| 1867529 | 0×2a980e | "There is a lad here, which hath five barley l... | test |
| 1867530 | 0×316b80 | When you buy the last 2 tickets remaining for ... | test |
| 1867531 | 0×29d0cb | I swear all this hard work gone pay off one da... | test |
| 1867532 | 0×2a6a4f | @Parcel2Go no card left when I wasn't in so I ... | test |

411972 rows × 3 columns

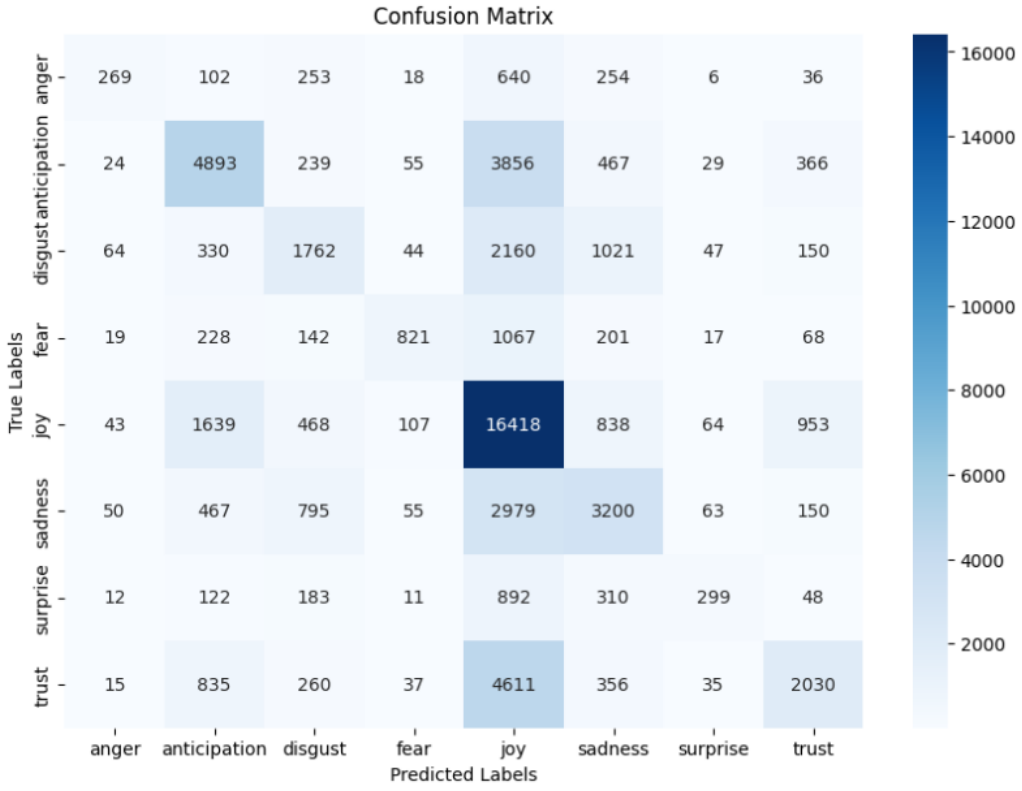
## 4.2 Model Selection

I have tried different approaches: TF-IDF with Logistic Regression, Word2Vec with CNN, and LLMs such as RoBERTa. The first approach gives a better result than the second one. In my conjecture, the third will have a much better result, since it will best fit this kind of task. However, since its computational cost is large and some setup is not done properly, regrettably I was not able to get a result from this approach.

## 4.3 Evaluation

In this section, I will provide the evaluation of the model using TF-IDF with Logistic Regression. The following figures are the classification result and confusion matrix:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| anger | 0.54 | 0.17 | 0.26 | 1578 |
| anticipation | 0.57 | 0.49 | 0.53 | 9929 |
| disgust | 0.43 | 0.32 | 0.36 | 5578 |
| fear | 0.72 | 0.32 | 0.44 | 2563 |
| joy | 0.50 | 0.80 | 0.62 | 20530 |
| sadness | 0.48 | 0.41 | 0.44 | 7759 |
| surprise | 0.53 | 0.16 | 0.25 | 1877 |
| trust | 0.53 | 0.25 | 0.34 | 8179 |
| accuracy | | | 0.51 | 57993 |
| macro avg | 0.54 | 0.36 | 0.40 | 57993 |
| weighted avg | 0.52 | 0.51 | 0.49 | 57993 |




# 5 Discussion and Future Works

From the confusion matrix above, we see that the training dataset is highly imbalanced. Therefore, choosing a model that can deal with an imbalanced dataset or performing a sampling technique to obtain a balanced subset is essential.

Furthermore, we conjecture that this task does not need to train a model by ourselves. It is possible that there are pre-trained models which are suitable for identifying emotions from text. This may save some time to enable us to perform more advanced preprocessing.