MSML641 - Natural Language Processing

Association of Human Emotions using Sentiment Analysis

Isha Asalla, Martina De Luca, Abhishek Kotcharlakota

May 17th, 2022

Isha Asalla
Martina De Luca
Abhishek Kotcharlakota

Project Report
MSML641: Natural Language Processing

2022-05-17

1 Introduction

Humans exhibit a wide range of emotions while responding to stimuli and information. According to Plutchik's wheel of emotions, these myriads of emotions can be broadly classified under eight kinds – joy, anger, fear, sadness, surprise, trust, disgust, anticipation. For the course of our project, due to the limitations of data, the number of classes have been reduced to primary five – joy, anger, surprise, fear, sadness or six including disgust.

Social Networking is a platform for many of us to respond to events happening around the world, share and comment on opinions. Platforms like Twitter, Facebook and Reddit witness millions of such comments every day. These comments carry forward users' emotions. Hence this project focuses on capitalizing such information to understand the patterns of association in human emotions.

2 Problem Statement

The primary objective of this project is to utilize a random set of texts using Social Media APIs to draw the association probabilities between two or more kinds of emotions. This involves applying Sentiment Analysis on the data for recognizing the emotion exhibited. Statistics and Machine Learning techniques are then used on the labeled emotions to identify the association rules. The results denote the probabilities by which one type of sentiment is likely to produce other kinds.

3 Data Description

The data for this project is a large set of comments taken from random subreddits using the Reddit API. The dataset contains a list of multiple comments for a group of users. The dataset comprises of approximately 300,000 comments distributed over 500 users.

3.1 Data Processing

The comments taken from Reddit users are usually long and contain multiple sentences. Each of these sentences can contain multiple emotions as the commenter shifts his wording or subject or language of what he/she is planning to communicate. Hence, each comment must be split by sentences to later apply sentiment analysis on each of them to cover all the emotions acted on by the user. Splitting the comments is also important since sentiment analysis algorithm that will be used is trained on tweets which are also shorter than the Reddit comments.

Further, like any other social media, most accounts use emoticons to be more specific or to make their responses short or attractive. These emoticons are essentially represented as groups of special characters in the dataset collected and therefore needed to be removed as a part of data processing.

Finally, the text from the comments contained many redundant characters. Formatting the data involved removing url links, random punctuation marks, capitalizations and extra white spaces.

4 Features

Initially, the data extracted from the API consisted of the following columns - commented ,author, body, created_utc, distinguished, edited, is_submitter, link_id, parent_id, saved score, stickied, subreddit_id. These features provide information about the user's ID, whether the comment has been edited, the comment date, the subreddit under which the comment has been posted etc. However, for this project 'body'(comment text) , 'author'(user id) ,'create_utc'(comment date) were majorly utilized.

5 Method

5.1 Sentiment Analysis

Before we understand the dependency of emotions, it's a prerequisite to recognize the primary emotions exhibited by a specific user. We applied two different approaches for recognizing multiple emotions from the reddit comments.

5.1.1 Soft Labelling

We used a twitter dataset with approximately 200,000 tweets gold labelled with 6 emotions. After removing punctuation, spaces, capitalization, truncating to maximum size, the vocabulary is tokenized into numerical codes. An LSTM model is then trained on the data to identify emotions in text. This model is used to soft label Reddit data. Since Twitter and Reddit posts' lengths are different, the Reddit comments needed to be split into sentences. Then the model is used on each sentence to obtain the emotion. Finally, the comment emotion is set to the one that appeared the most.

5.1.2 Text2Emotion

This is a pre-trained Python package for emotion detection with 5 emotions. This package processes any textual message and recognizes the emotions embedded in it. This is not a time-series recognition. Two approaches were taken while applying this algorithm:

- Considering the emotion with the maximum value.
- Considering the probability of each emotion and adding up the probabilities.

5.1.3 Grouping Users' Emotions

The users' emotions were recognized with the afore-mentioned algorithms. Each user is assigned counts for every kind of emotion that has been recognized from the corresponding account tweets, an example is shown in Table 1. These counts are then used to calculate the individual emotion probabilities for each emotion per user, as shown in Table 2, which are then incorporated into the association model.

User ID	Anger	Fear	Joy	Sadness	Surprise
spoopydoopyjoopy	37	117	346	114	163
BadKarma-18	8	90	141	51	68
Peoerson	15	119	135	47	75
Doomed	32	404	247	154	162
lobester250	1	1	832	3	1

Table 1: Count of posts by emotion for each user from Text2Emotion - Maximum algorithm.

User ID	Anger	Fear	Joy	Sadness	Surprise
spoopydoopyjoopy	0.047619	0.150579	0.445302	0.146718	0.209781
BadKarma-18	0.022346	0.251397	0.393855	0.142458	0.189944
Peoerson	0.038363	0.304348	0.345269	0.120205	0.191816
Doomed	0.032032	0.404404	0.247247	0.154154	0.162162
lobester250	0.001193	0.001193	0.992840	0.003580	0.001193

Table 2: Probabilities by emotion for each user from Text2Emotion - Maximum algorithm.

5.2 Association Rules

Association rule-mining is a data mining approach used to explore and interpret large transactional datasets to identify unique patterns and rules [1]. These datasets contain a large amount of transactions where each of them include a list of items. It is important to note that these transactions only indicate which items are included, not the amount in each one.

For our project, we will consider our transactions as each of the users and the items as the emotions. Since most users will have at least one comment of each emotion, this will generate a few challenges when applying the existing algorithms, because for our project it is significant the value of either the count or probability for each emotion that we generated by the sentiment analysis algorithm.

Some important metrics to consider for association rule algorithms are described below [2].

- Support count(E1): is the number of transactions containing Emotion 1.
- Support(E1): is the support count divided by the number of total transactions.

$$Support(E1) = \frac{\text{Number of transactions containing Emotion 1}}{\text{Total number of transactions}}$$

- Confidence(E1 → E2): indicates out of all the times emotion 1 appears, how many times is also emotion 2 present.

$$Confidence(E1 \rightarrow E2) = \frac{\text{Number of transactions containing E1 and E2}}{\text{Number of transactions containing E1}}$$

- Lift(E1 → E2): indicates how often the emotions appear together in comparison to what is expected by chance.

$$Lift(E1 \rightarrow E2) = \frac{Confidence(E1 \rightarrow E2)}{Support(E2)}$$

We applied two association rule algorithms to identify correlation between the emotions exhibited in the posts. We generated the association rules from the results of the three sentiment analysis approaches. The results shown in the next sections will be regarding the Text2Emotion - Maximum approach.

5.2.1 Apriori Algorithm

The Apriori algorithm focuses on the most frequent items (in this case emotions) to identify rules between them. The approach of this algorithm is to take the most frequent items by highest support count and create combinations between them, from all these possible combinations it generates the rules by maintaining the most frequent combinations, measured by using metrics such as support, confidence and lift as described above [2].

As mentioned previously, since the structure of the dataset is not exactly as what the Apriori algorithm expects, if we were to apply it as is, most users would have a list of all emotions, and the algorithm would not provide any useful insights.

Therefore, we decided to implement a custom version of the algorithm by following the same process but multiplying by probabilities in each step of the support count calculation, instead of just performing the simple count. We created all possible combinations of up to three emotions and generated the association rules. Finally, we kept those which had the highest value of confidence, ordering the results by lift. This means that we are keeping those rules which appear more frequently, but we see the results in order of which rules are appearing more than they would by chance.

5.2.2 Frequent Pattern Growth Algorithm

The Frequent Pattern Growth algorithm (also known as FPGrowth) was created as an improvement to the Apriori algorithm, to improve efficiency. As a result it builds a tree with associations in branches based on the transactions from the dataset, using the same metrics described previously to identify the most frequent emotion rules [3].

Similar to the challenge presented above for the application of the Apriori algorithm, the direct application of the FPGrowth algorithm to our current dataset will most likely result in one unique branch with all emotions. This is because all users would have at least one post on each emotion and this would be the pattern detected.

For this case, we decided instead to try a different approach and transform the data set to suit the default algorithm. For this purpose, we created levels of emotions to simulate a difference between the items of each user (concatenating the emotion name and the level tag). The levels of emotions defined are described below. The tag ranges are not evenly distributed because lower values are more common in the dataset, therefore distributing the ranges equally would result in most items tagged as low or mid-low and very few in the other categories.

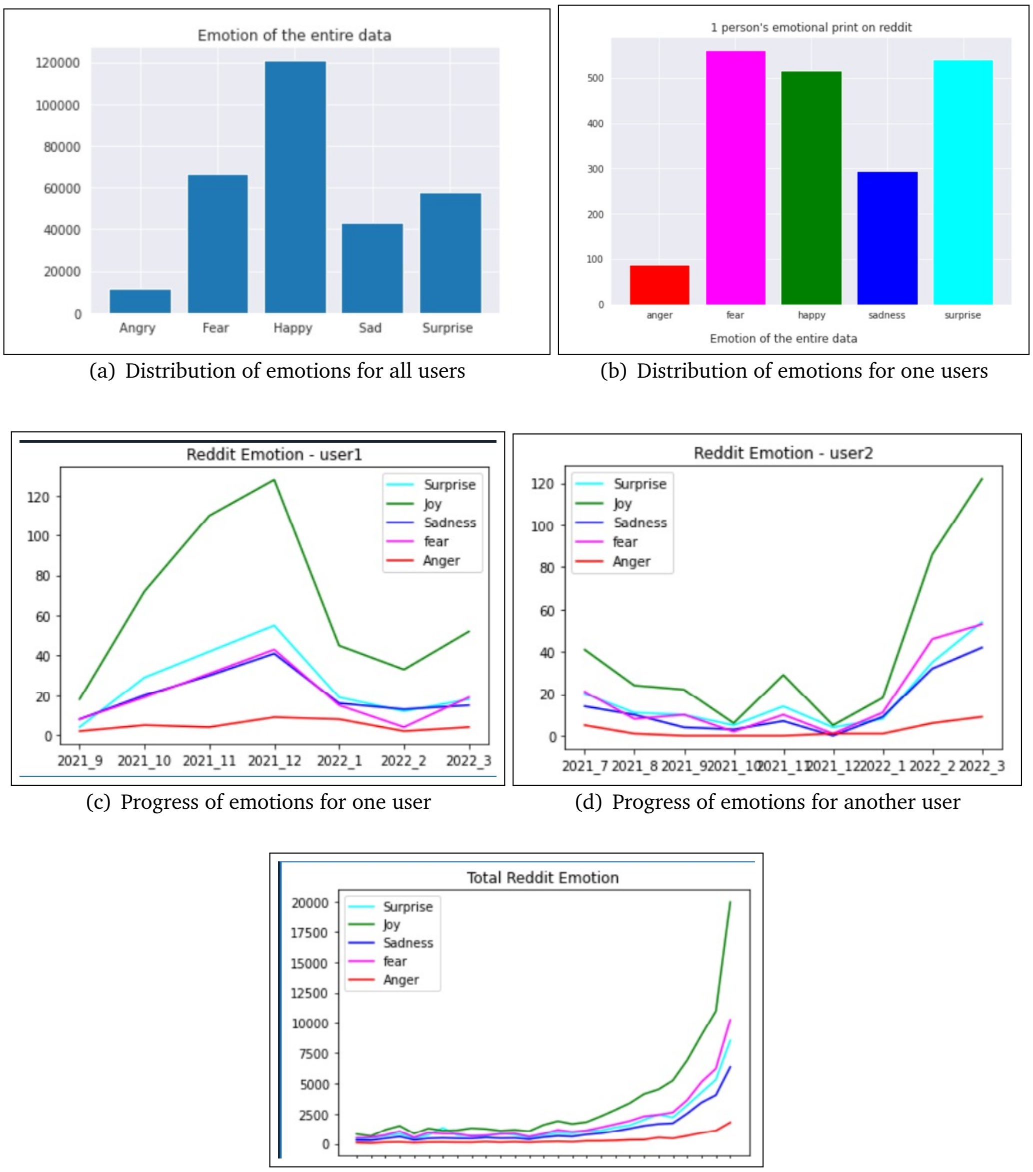
- Low (L): 0 - 0.19
- MidLow (ML): 0.2 - 0.29
- Mid (M): 0.3 - 0.39
- MidHigh (MH): 0.4 - 0.59
- High (H): 0.6 - 1

After updating the dataset, each user would contain a list of items indicating for each emotion the level. From this dataset we were able to apply the predefined FPGrowth function from pyfpgrowth library [4] and incorporated association rules with highest confidence, ordered by lift.

6 Results

6.1 Sentiment Analysis

As a result of the sentiment analysis processes we could see an overview of how the emotions are displayed for this specific data set. Below we present the most relevant graphs that show how these emotions are distributed for the Text2Emotion - Maximum approach. First, $Plot\ a$ and $Plot\ e$ show how the emotions are represented and distributed for all users. Then, $Plot\ b$, $Plot\ c$ and $Plot\ d$ display the same distributions for individual random users over several months to see what is the common emotion for a user.



(a) Distribution of emotions for all users

(b) Distribution of emotions for one users

(c) Progress of emotions for one user

(d) Progress of emotions for another user

(e) Progress of emotions for all users

6.2 Association Rules

The following are the results obtained from applying the two association rule algorithms to the results of the sentiment analysis of the Reddit comments.

After applying the Apriori algorithm, we got the following rules as a result.

Rule	Confidence(%)	Lift(%)
Sadness, Surprise → Fear	21.97	20.27
Surprise → Fear	21.13	19.49
Surprise → Joy	38.15	17.44
Sadness → Joy	36.18	16.54
Sadness, Surprise → Joy	35.28	16.13
Fear, Surprise → Joy	33.87	15.48
Fear → Joy	33.47	15.30
Fear, Sadness → Joy	32.03	14.64

After applying the FPGrowth algorithm, we got the following rules as a result.

Rule	Confidence(%)	Lift(%)
Anger-Low, Joy-MidHigh → Sadness-Low	96.28	107.18
Anger-Low, Fear-MidLow → Sadness-Low	92.77	103.28
Anger-Low, Joy-Mid → Sadness-Low	91.67	102.05
Anger-Low, Surprise-MidLow → Sadness-Low	91.49	101.85
Anger-Low, Fear-Low → Sadness-Low	90.73	101.01
Joy-Mid, Sadness-Low → Anger-Low	100	100.99
Joy-MidHigh, Sadness-Low → Anger-Low	100	100.99
Sadness-Low, Suprise-MidLow → Anger-Low	99.42	100.40
Fear-MidLow, Sadness-Low → Anger-Low	99.35	100.34
Fear-Low, Sadness-Low → Anger-Low	99.16	100.14

From these two results we can observe that the outcomes are very different between the two algorithms, this can be explained by the fact that the approach and data set used for both are different.

For the Apriori algorithm, we are focusing on the emotions that appear more frequently and we can observe that reflected on the rules, where the emotions that are in most of the rules are joy and fear, which are the most frequent emotions in the entire dataset as seen on the results of the sentiment analysis. Another thing to note for the Apriori algorithm is the low confidence and lift values, this is because the support count is calculated by adding the probabilities of the emotions for each user, and when creating the combinations of the emotions these values are multiplied, resulting in smaller values.

On the other hand, for the FPGrowth algorithm, we are still focusing on the most frequent items, but in this case that is the most frequent level of each emotion. So even though anger is on of the less popular emotions, many users seem to have a low level of anger and that is frequent in the rules of the outcome.

7 Future Direction

We plan to allocate a separate classification for a neutral emotion as most comments need not be deciphered as expressing any specific emotion. More behavioral traits such as kindness, humour, selfishness etc can be included in our model and find specific associations. Time factor has not been considered in this project but if applied more accurate results may be obtained.

Furthermore, other association rule algorithms can be used and the current ones improved to go into more detail on certain algorithms and attempt other transformations to the data to understand if this would give more insightful results.

8 Conclusion

Before stating the conclusions, we must note that these results represent a subset of Reddit data. In addition, we can't affirm that these represent the actual human emotions, since what is posted on social media doesn't necessarily need to be the same as the emotion the user is going through and not all emotions are posted equally by all individuals.

Results from each algorithm of sentiment analysis vary in the identifications made of the emotions for each post. Presence of neutral emotion or no classification for a given text has an important impact on the results, so we identified this as an important factor to include in future work. Using the sentiment analysis provided results to show how the emotions were distributed between them and though time.

From the application of the association rules we can see that both algorithms provide valuable insights. On the one hand , the Apriori algorithm focuses on the most frequent emotions considering the number of appearances, whereas FPGrowth algorithm helps identify common levels of emotions between the users. As a result from both these algorithms, and taking into account the considerations mentioned above, we could say that there is a high presence of joy and fear, but the difference is that joy usually is present on high levels and fear on lower levels. Similar to fear, we can identify that surprise is frequently present but with low levels in the users. We can also observe that both anger and sadness are present in a high number of users in a low level of emotion.

As a result, this project has been a great learning experience where we were able to apply multiple Natural Language Processing models, compare them and then use the results with association rules to get additional insights on the data.

References

[1] Abdulhamit Subasi. Association Rules - Chapter 3 - Machine learning techniques, 2020. URL `https://www.sciencedirect.com/topics/computer-science/association-rules`.

[2] Amit Ranjan. Apriori Algorithm in Association Rule Learning, December 2020. URL `https://medium.com/analytics-vidhya/apriori-algorithm-in-association-rule-learning-9287fe17e944#:~:text=What%20is%20Apriori%20Algorithm%3F,which%20are%20frequently%20visited%20etc`.

[3] Joos Korstanje. The FP Growth algorithm, September 2021. URL `https://towardsdatascience.com/the-fp-growth-algorithm-1ffa20e839b8`.

[4] Evan Dempsey. Python Documentation. URL `https://pypi.org/project/pyfpgrowth/`.