Myers-Briggs Type Indicator prediction with LSTM

Omer Sarid
https://github.com/saridgnr/NLP

1 Introduction

The Myers Briggs Type Indicator (or MBTI for short) is a personality type system that divides everyone into 16 distinct personality types across 4 axis:

- Introversion (I) - Extroversion (E)
- Intuition (N) - Sensing (S)
- Thinking (T) - Feeling (F)
- Judging (J) - Perceiving (P)

More than 2.5 million people a year take the test globally. Some 89 of the Fortune 100 companies in the US use it as a human resources tool, and it is used widely across Australia and Asia. Its history stretches back over seven decades.

Although popular in the business sector, the MBTI exhibits significant scientific deficiencies, notably including poor validity, poor reliability (giving different results for the same person on different occasions), measuring categories that are not independent, and not being comprehensive . The four scales used in the MBTI have some correlation with four of the Big Five personality traits, which are a more commonly accepted framework.

In this project I used bidirectional LSTM and attention to try and predict a person's MBTI from a given text.

2 Data Set

Reddit is an American social news aggregation, web content rating, and discussion website. Registered members submit content to the site such as links, text posts, and images, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "subreddits", which

cover a variety of topics including news, science, movies, video games, music, books, fitness, food, and image-sharing.
As of March 2019, Reddit had 542 million monthly visitors (234 million unique users), ranking as the #6 most visited website in U.S. and #21 in the world, according to Alexa Internet, with 53.9 % of its user base coming from the United States, followed by the United Kingdom at 8.2% and Canada at 6.3%.

Along with the many subreddits there are in Reddit, there is also a subreddit for each distinct personality type. I used Reddit API to query the relevant

Figure 1: Every MBTI's personality type/subreddit

subreddits for their top posts and their comments to create the data set I used for this project. The data set is in TSV format,
Each line in the file represents:

- A Post: Personality, Upvotes, Title, Content
- A Comment: Personality, Upvotes, Content

The data set contains over 110k posts and comments from all the MBTI subreddits combined. There is also a smaller data set that contains 12k posts and comments only, that was used for experimentation.

3 Architecture

For this project I used Document classification with LSTM and Attention.

4 Experimentation

`https://github.com/saridgnr/redditmbti/tree/master/visualize`

Most of the experiments results I did are specified in the link above, When I started this project I picked arbitrary parameters for the network that appeared to be working and started from there(These parameters are specified in exp2). The epoch size I used for all of the documented experiments is the length of the entire training data, I found it easier to measure progression that way. Each experiment ran for 7 epochs

At first I wanted to try to see how does the hidden size effects the quality of the network, I found out that for lower hidden sizes the loss seems to decrease in a linear way but the precision, recall and fscore did not improve that much, That means that the network over fitted to the training data, Higher hidden sizes to an extent made the quality of network appear to be better but also harder to train.
After picking an hidden size that appeared to be good I wanted to see how does the embedding size affects the network, When lowering the embedding size the loss did not appear to improve at all, and increasing it seem to improve the precision recall and fscore. I also tried using much larger embedding and hidden size but even though the loss was improving the evaluation was not as good as the medium sizes.
At last I tried to change the number of LSTM layers, I found out that the optimum for my project is two.
The best experiments were exp2 and exp4,
I picked the parameters of exp2 and tried it on the entire data set which contains 110k posts and comments and ran it for 13 epochs.

5 Final Results

5.1 Exact MBTI

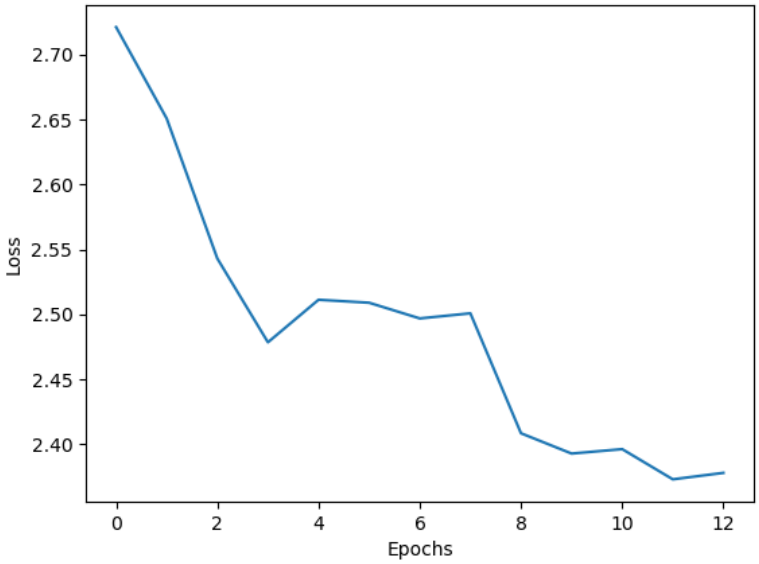

Figure 2: The loss for epoch plot

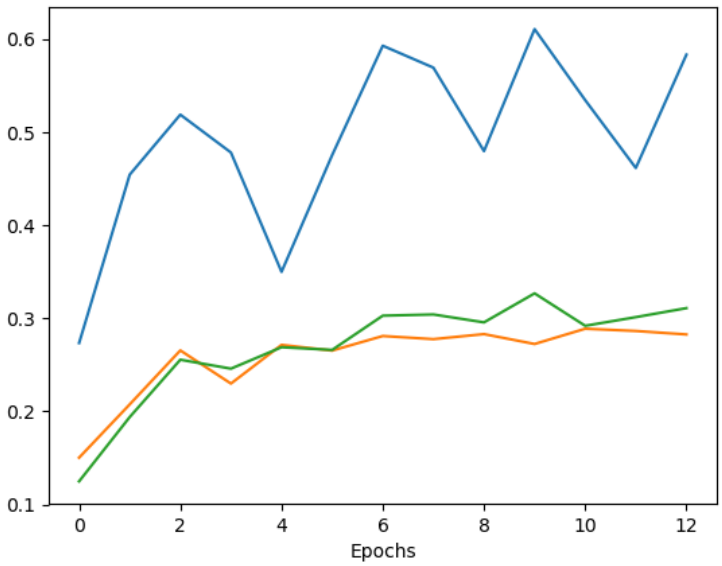

Figure 3: precision/recall/fscore for epoch plot

5.2 Extrovert-Introvert axis

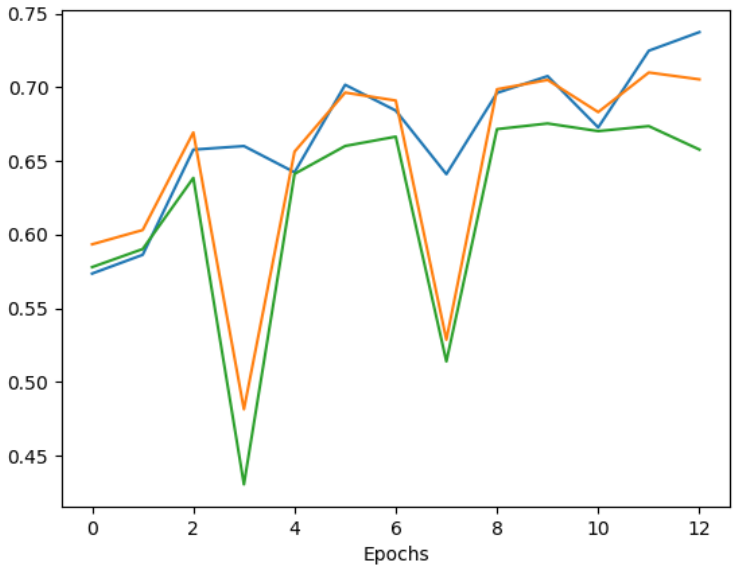

Figure 4: precision/recall/fscore for epoch plot

5.3 Sensing-Intuition axis

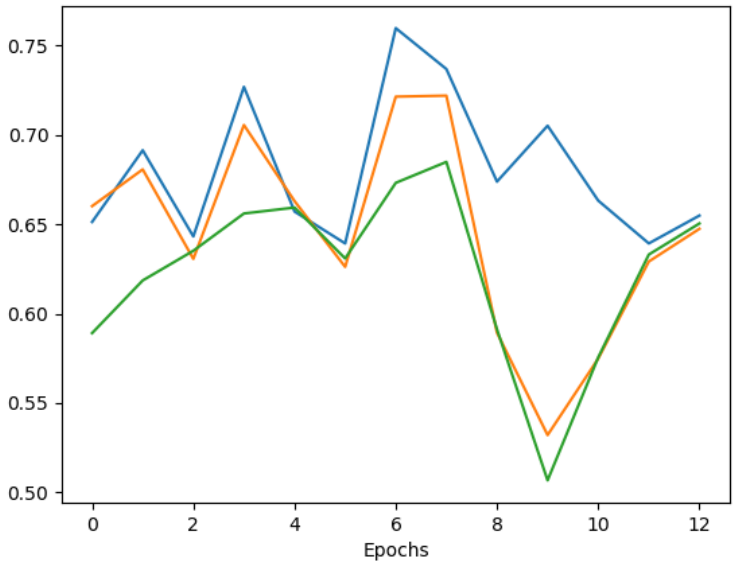

Figure 5: precision/recall/fscore for epoch plot

5.4 Thinking-Feeling axis

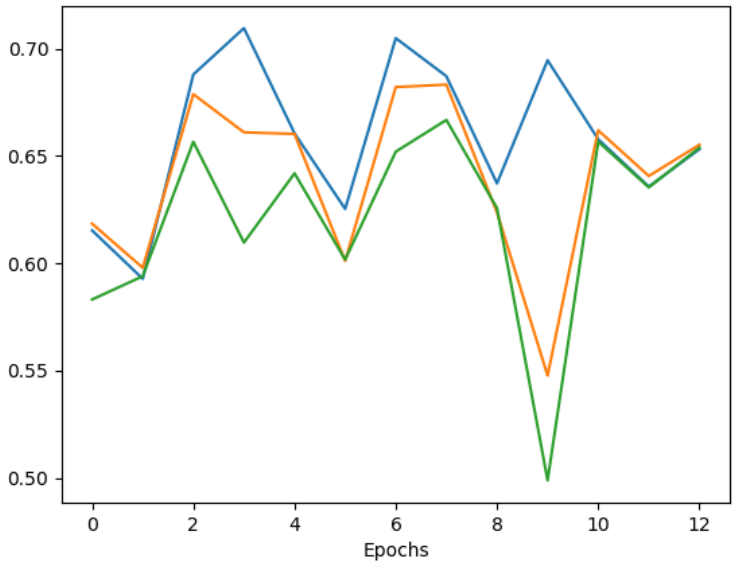

Figure 6: precision/recall/fscore for epoch plot

5.5 Judging-Perception axis

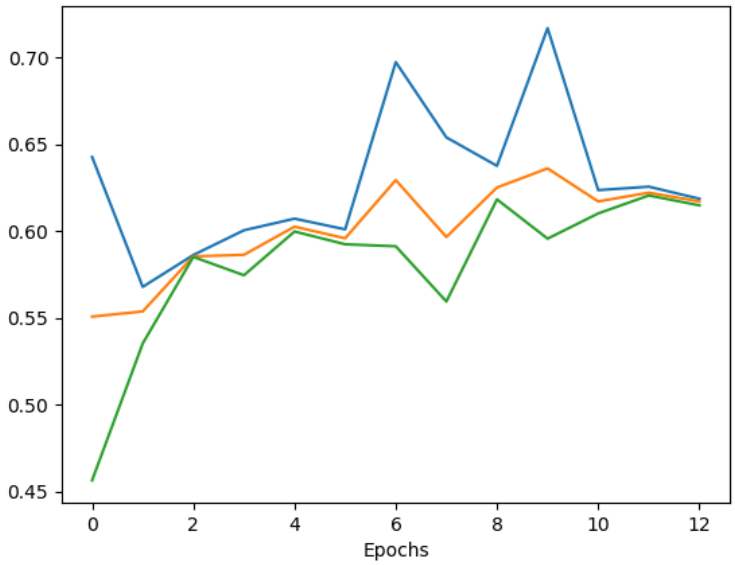

Figure 7: precision/recall/fscore for epoch plot

6 Conclusions

- The subreddits of MBTI are noisy: people are posting and commenting mostly in their own MBTI subreddits but not only, which causes a lot of noise in the data set.
- Some axes are easier to predict than others
- MBTI is a poor personality type indicator: MBTI categories are criticized for not being independent(among other things) which makes predicting a person's MBTI a tough task(and not only for a NN)
- Bigger data sets may cause over fitting to the test data.
- While increasing the number of parameters may improve the network's results per epoch it also increases the training time.