DSC96 Final Project

Prepared by: Scramble Eggs(Jun Linwu & Man-Fang-Liang)

Motivation:

This project would be focused on exploring the behaviors, preferences, and others' perceptions on the selected MBTI types, ENTJ and INTJ. The motivation behind this project is our common interest in developing a further understanding of our own MBTI type.

Research Question

- What are the most typical impressions/words associated with ENTJ and INTJ, are they consistent with the description from authority websites?
- Which MBTI type receives more positive feedback during human interaction, ENTJ or INTJ?

Data

- The MBTI-Types & Enneagram Texts dataset, as a secondary data source, was collected from Kaggle (https://www.kaggle.com/yamaerenay/mbtitypes). It was primarily collected from web-scraping the Reddit community.
- Cleaning: Before the analysis, we extract the content with unrelated labels and forum, ie. Romance, MBTI forums beside the targeted types. We create two separate datasets, data_INTJ and data_ENTJ. To increase human readability, we divided one *texts* cell into multiple cells (one response per cell). Then, the duplicate and irrelevant content is removed (ie. other website link, failed quotation, testing on send and reply, etc.) from the *texts* column using Google sheet.
  https://docs.google.com/spreadsheets/d/1XRnP3EuqBSOsH5YAeTYCbC4tp_XdY89A5Xr9AlkwYWI/edit?usp=sharing
- data_INTJ and data_ENTJ have similar structures. Each consists of 5 columns, 291 rows for data_INTJ and 49 rows for data_ENTJ, 340 rows in total.
  index: identification number of the post
  title: title of the post
  texts: question or the post content
  labels: category of the post
  forums: the community where the post was sent with relevant ideas

Research Method

- An open-coding analysis would be performed using Excel. Each author would work on about 22% (the first 74 data) of the data_intj individually to assign a sentiment score (1, 0, -1; positive, neutral, negative accordingly) and a typical impression for each text.

- The codes would be adjusted until Cohen's kappa value reaches over 0.80. We have come to final agreement through discussing some of the ambiguous cases and set the standard code for the rest evaluation of the typical impression.

  Standard code for typical impression:
  - N/A: excluded for analysis
  - Intellectual: smart, wit
  - Less sensitive: less emotional, more objective
  - Fairly sensitive: emotional as normal
  - Dominant: commanding, controlling
  - Humorous: funny
  - Perseverance: consistent working-hard
  - Mind-Strong: independent, discipline
  - Challenging:
  - Honest: straightforward in judging
  - Fake: act contradicted to what they said
  - Quiet
  - Skeptical: unlikely to trust easily
  - Non-expressive: less communicative
  - Procrastinated:
  - Introspective: introverted
  - Prospective: have broader perspective

- Once the combined codebook is implemented, we make a second dataset containing the overall percentage of the relevant behaviors and perceptions(positive, neutral, and negative) for each MBTI type.
- An additional sentiment analysis would be performed using NLP using Python importing nltk and babypandas packages. The following result would serve as a sanity check that whether the result of the open-coding analysis is consistent with the result of NLP.

Results

Open-Coding:

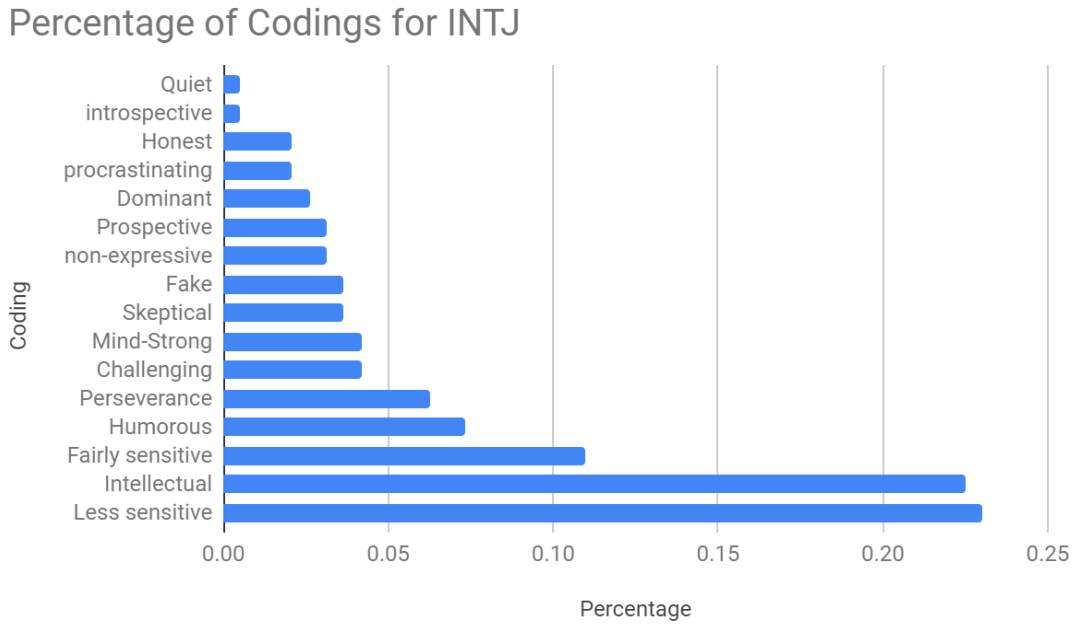
- Kappa for sentiment: 0.8029585799 > 0.80
- INTJ (n=291)
  Negative: 38 (0.131)
  Neutral: 178 (0.612)
  Positive: 75 (0.258)

- ENTJ (n=49)
  Negative: 2 (0.041)

Neutral: 35 (0.714)
Positive: 12 (0.245)

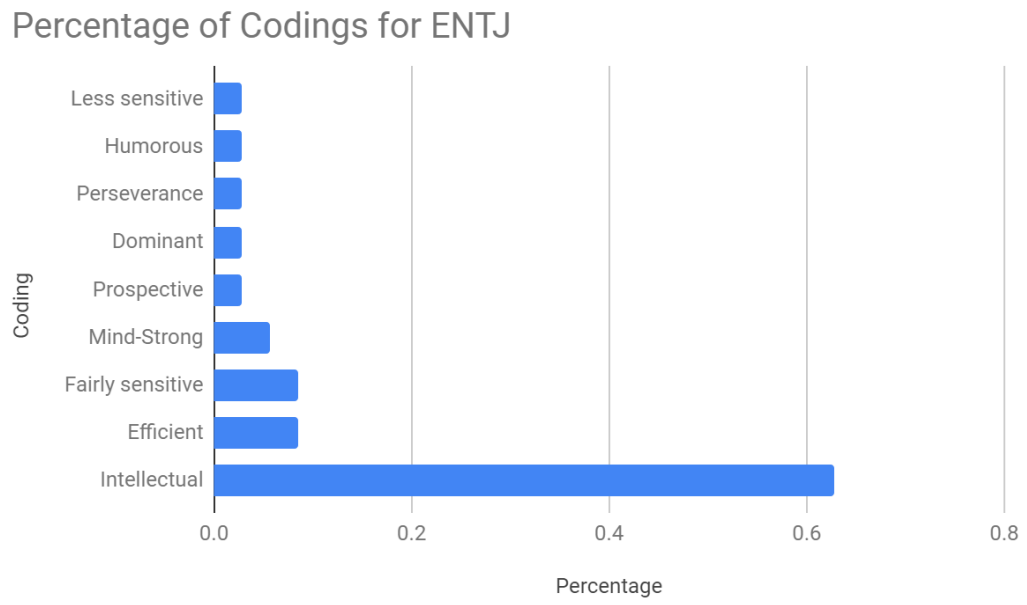
- Kappa for impression: 0.8068587106 > 0.80
- INTJ (n=191 excluded N/A of 113)

Coding	Count	Percentage
N/A	113	NA
Intellectual	43	0.225
Less sensitive	44	0.230
Fairly sensitive	21	0.110
Humorous	14	0.073
Perseverance	12	0.063
Dominant	5	0.026
Mind-Strong	8	0.042
Challenging	8	0.042
Honest	4	0.021
Fake	7	0.037
Quiet	1	0.005
Skeptical	7	0.037
See things glob	6	0.031
introspective	1	0.005
non-expressive	6	0.031
procrastinating	4	0.021
	191	1



- ENTJ (n=35 excluded N/A of 20)

Coding	Count	Percentage
N/A	20	NA
Intellectual	22	0.629
Less sensitive	1	0.029
Fairly sensitive	3	0.086
Humorous	1	0.029
Perseverance	1	0.029
Dominant	1	0.029
Mind-Strong	2	0.057
Efficient	3	0.086
Prospective	1	0.029
	35	1



Average Sentiment Score by Python: https://github.com/juxx8/dsc96-final_project

- INTJ
  - Negative: 0.070
  - Neutral: 0.773
  - Positive:0.154
  - Compound: 0.393

- ENTJ
    - Negative: 0.065
    - Neutral: 0.775
    - Positive: 0.156
    - Compound: 0.416

Implication & Conclusion

- In both approaches to analyzing sentiment, ENTJ receives less negative feedback from their interaction with others than INTJ does.
- However, a contradicting result was found on positive feedback that INTJ receives more positive feedback when analyzing sentiment using open-coding approach yet ENTJ receives more positive feedback when using NLP approach.
- In both approaches, people tend to feel neutral when interacting with INTJ and ENTJ despite what their own MBTI types are, implying that most people recognize, accept, and even enjoy the differences between people.
- Intellectual is the most typical impression for ENTJ, while Less Sensitive is the most typical impression for INTJ.
- Both INTJ and ENTJ tend to be recognized as intellectual, which is consistent with the description of “T”, which denotes Thinking.
- We realized that there is a much larger sample size for INTJ than ENTJ, one potential reason might be that the extrovert would focus on more in-person interaction while an introvert tends to participate more on online conversations.

Challenge Faced

- We generally spend most of our time cleaning the data because there are undetected irrelevant posts such as testing reply posts that could not be recognized and removed without manual check.