Gender Stereotype of Rationality: A NLP Analysis Based on Reddit Posts and MBTI Types

Iris Shi
1778676

Jikun Shen
1833847

Abstract

In this paper, we verified a common gender stereotype: men are more rational than women. The feeling and thinking Reddit post dataset was used to train and test models and the gender Reddit post dataset was used to predict "Feeling" and "Thinking" classes based on post texts. SGDClassifier model is selected as the classifier because it has the highest accuracy (0.84) compared with Logistic Regression, Random Forest Classifier and LSTM models. We found that males are more likely to be associated with thinking, while females have a more balanced distribution between thinking and feeling, with a slightly higher proportion leaning towards thinking. The finding contradicts the gender stereotype that men are more rational and women are more emotional.

1 Introduction

Are men more rational than women? The notion that men are more rational than women is a long-standing and widely accepted gender stereotype. This belief has been ingrained in various aspects of Western thought and societal norms, shaping the way individuals perceive and reason about different situations(Jones, 2004).

The Myers-Briggs Type Indicator (MBTI) is a widely used personality test that categorizes individuals into 16 different personality types based on four dichotomies: Extraversion vs. Introversion, Sensing vs. Intuition, Thinking vs. Feeling, and Judging vs. Perceiving. The dichotomy of Thinking vs. Feeling describes the preference for how people make decisions, by relying on logic or emotions towards people and special circumstances, which is related to rationality judgement. (Stajner and Yenikent, 2021b; Briggs-Myers and Myers, 1995)

How to detect "Thinking" and "Feeling" from texts? The use of questionnaire-based personality detection presents several limitations. It not only necessitates training human assessors, which can be resource-intensive and subject to inter-rater variability, but is also susceptible to social desirability bias.(Heine et al., 2002) The automatic detection of MBTI gained popularity in recent years and Twitter data is most common to use to train a classifier.(Plank and Hovy, 2015). All those studies, despite using large training datasets and various features, barely managed to outperform the majority-class baseline, and even that only in some of the four MBTI dimensions.(Stajner and Yenikent, 2021a)

The central research question of this paper is: Are men more rational than women? We also want to figure out which model performs best in detecting "Thinking" and "Feeling" from textual data and find the best parameter settings. After text cleaning work for the Reddit post dataset, we used Logistic Regression, Random Forest Classifier, SGDClassifier, and LSTM model to classify "Thinking" and "Feeling" and used a pipeline that includes a TF-IDF Vectorizer to find the best parameter settings. Then, we selected the best performance model to predict "Thinking" and "Feeling" for a Reddit post dataset which has gender labels to answer the research question.

2 Related Work

Several studies have employed machine learning methods to explore gender biases and stereotypes in various contexts, such as literature, artificial intelligence and online communities to investigate gender-related patterns and biases in different domains.

For example, a massive machine learning study demonstrated gender stereotyping and sexist language in literature by analyzing 3.5 million books and identifying fundamental differences in the written language used to describe men and women (Hoyle et al., 2019). Another study focused on

exposing implicit biases and stereotypes in human and artificial intelligence, highlighting the permeation of human biases and stereotypes in machine learning algorithms trained on natural language data (Marinucci et al., 2023). Additionally, research has been conducted to investigate male and female users' differences in online technology communities using supervised machine learning methods, which revealed significant differences in user behaviour based on gender (Sun et al., 2020).

Automatic detection of MBTI is very popular in recent years and Twitter data is most common to use to train a classifier. Plank and Hovy (2015) outperformed the majority-class baseline only on the IE and TF dimensions, achieving the accuracy of 72.5% and 61.5% on those binary tasks. LSTM was applied to MBTI detection and got 0.99 accuracy(MATHUR, 2021).

3 Data

Data comes from a social media named Reddit and is organized into 2 datasets, named "gender" and "feeling and thinking". Data preprocessing procedures and an overview of these two databases are introduced below.

3.1 Data Preprocessing

Text data from social media usually contains errors and messy elements, which will affect the performance of the training model. Before analyzing the data, several steps were undertaken to reduce the noise and overlapped information of these two datasets.

3.1.1 Noise Reduction

Disordered numbers, emojis, MBTI, HTML tags, punctuation, stopwords, special characters, and words and letters that are repeated multiple times next to each other were removed to minimize noise."nltk" package in Python was used to remove stopwords. In addition, the Spello library in Python was used for spelling correction.

3.1.2 Lemmatization

Lemmatization simplifies words to their root forms, which is useful for information retrieval, where word variants are treated as equivalent. Considering the time cost and previous experience, we did lemmatization before the model training procedure and spacy package was used.

Table 1 shows a text example after data preprocessing.

Figure 1: Word cloud of "Feeling" class.

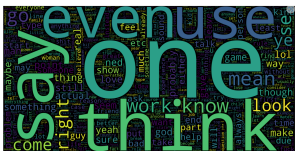

Figure 2: Word cloud of "Thinking" class.

3.2 Feeling and Thinking Reddit Post Dataset

The feeling and thinking Reddit post dataset has three columns: author ID, post, and feeling. The feeling column indicates whether each post is categorized as rational or emotional, with "1" representing "Thinking" and "0" representing "Feeling". This dataset has 1366 posts labelled as "Thinking" and 701 posts labelled as "Feeling" after grouping data by author ID, which is unbalanced. Table 2 shows that the average post length is 90132 and the average word length is 5.3.

We did word cloud analysis of "Feeling" and "Thinking" classes based on the post corpus after data preprocessing. Figure 1 and Figure 2 illustrate that "one", "think" and "say" appear frequently in both "Thinking" and "Feeling" Classes. "Man" and "Woman occur in both figures. The font size of them is small, which guarantees our model will be slightly influenced by "Man" and "Woman".

3.3 Gender Reddit Post Dataset

The gender Reddit post dataset comprises author ID, post, and gender information. In the gender column, "1" represents "Female" and "0" represents "Male". This dataset has 1281 posts authored by females and 1120 posts authored by males after grouping data by author ID, which is approximately balanced. Table 2 shows that the average posts length is 90132 and the word length is 7.

4 Experimental Setup

The test size was set at 0.3. Considering the unbalanced data distribution, the "stratify" parameter was also set. A pipeline was used in scikit-learn that included a TF-IDF Vectorizer and a Classifier model. Four experimental models were explored. The GridSearchCV is employed to explore differ-

Original Post	After Preprocessing
Sending love and light They're both single so I don't understand why they have to hide their relationship.	send love light single understand hide relationship

Table 1: An example of a post after data preprocessing.

	Post Length(F&T)	Word Length(F&T)	Post Length(G)	Word Length(G)
Mean	90132	5.3	148842	4.4
Min	3970	1	7451	1
Max	4096354	5.8	4407965	7

Table 2: An example of post after data preprocessing. F&T means Feeling and Thinking Reddit post dataset. G means gender Reddit post dataset.

ent settings for TF-IDF Vectorizer's lowercase and max-features parameters.

4.1 Experimental Models

4.1.1 Logistic Regression

Logistic Regression is a linear model used for binary classification, which is commonly used for classification. It models the probability of the default class (0 or 1) based on input features.

4.1.2 Random Forest

Random Forest is an ensemble learning method based on decision trees. It builds multiple decision trees and merges their predictions to improve accuracy and reduce overfitting.

4.1.3 SGDClassifier

SGDClassifier is a linear classifier that uses stochastic gradient descent for optimization. It's a flexible model capable of handling large datasets and is useful for various types of classification tasks.

4.1.4 LSTM

LSTM is a type of recurrent neural network (RNN) architecture designed to learn long-term dependencies on sequential data. In this paper, LSTM is designed for binary text classification tasks, and it incorporates pre-trained GloVe word embeddings to enhance the representation of words in the embedding layer. The LSTM layers capture sequential dependencies on the input data.

4.2 Parameters

For Logistic Regression, Random Forest Classifier, SGDClassifier and LSTM model, Lowercase and max-features parameters were explored. For the max-features parameters, we explored 5000 and 10000. We performed a grid search using our feeling and thinking data and combined labels to determine the best combination of values for our parameters, fitting 5 folds for each of 4 candidates, totalling 20 fits. The result shows that the best parameter settings are lowercase is "True" and the max-features parameter is 1000 for all 4 models.

Considering the time cost, the parameter settings of our LSTM model mainly refer to the settings of MATHUR (2021), which got 99.86% accuracy in classifying fake news on Twitter. The difference is that "trainable" was turned on in the Embedding layer in this paper.

5 Results

5.1 Model Performance

An overview of model performance on the feeling and thinking dataset can be found in Table 3. The table provides an overview of the evaluation of the predicted labels against the originals of each model in classifying the data.We report the weighted average of precision, recall, and F1 score due to unbalanced classes in our test data. These models include RandomForestClassifier, SGDClassifier, LogisticRegression, and an LSTM model.

The SGDClassifier shows balanced performance across both classes, which has the highest accuracy. The RandomForestClassifier has a high recall for the "Thinking" class(T) but a relatively low recall for the "Feeling" class(F), which means it's better at identifying one class over the other. The LogisticRegression model demonstrates good precision and recall but slightly lower accuracy compared to the SGDClassifier. The LSTM model has high recall for the "thinking" class(T) but poor performance for the "feeling" class(F), indicating a specific issue in recognizing instances of the "feeling"

class(F).

Table 4 indicates that for the "Thinking" class(T) and "Feeling" class(F), the baseline accuracy is 66% and 34%. The accuracy of the SGDClassifier is 82%, which is significantly higher than the baseline accuracy for both classes, which suggests that the SGDClassifier is performing well and learning meaningful patterns from the data. The accuracy of the RandomForestClassifier and LogisticRegression model is 75% and 81% separately, also higher than the baseline accuracy for both classes but lower than SGDClassifier. The accuracy of the LSTM model is 60%, which is lower than the accuracy of the other models. For the "feeling" class(F), the LSTM accuracy of 60% is higher than the baseline of 34%.

5.2 Prediction Result

The trained model is used to predict "Thinking" and "Feeling" between genders. Figure 3 suggests that, among males, 91.07% of predictions were classified as "Thinking" and 8.93% of predictions were classified as "Feeling". Among females, 72.29% of predictions were classified as "Thinking" and 27.71% of predictions were classified as "Feeling". There's a difference in the distribution of predicted categories between genders. For males, a higher percentage of predictions are classified as "Thinking", while for females, there's a more balanced distribution between "Thinking" and "Feeling," with a higher proportion leaning towards "Thinking" as well but not as pronounced as in males.

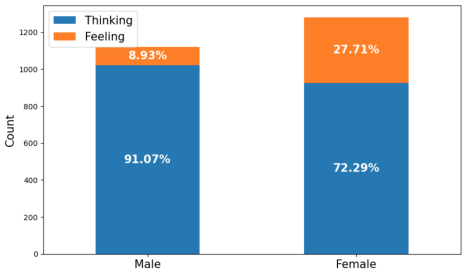

Figure 3: Comparison of Predicted "Thinking" and "Feeling" Classes Between Genders

6 Discussion and Conclusion

The results of the model performance section suggest that the SGDClassifier model performs best in predicting the "Thinking" and "Feeling" classes for the gender dataset compared with Logistic Regression, Random Forest Classifier and LSTM models. The accuracy of the SGDClassifier model can reach 0.82. The best parameter settings are lowercase "True" and the max-features parameter is 1000 for all 4 models. The accuracy is higher than in several studies. For example (VARMA, 2021) got 0.7 accuracy using an XGBClassifier to label "Feeling" and "Thinking" of Twitter posts.Plank and Hovy (2015) achieved an accuracy of 61.5% on the TF dimension. However, there are still models with higher accuracy in classifying "Feeling" and "Thinking" classes. For instance, MARQUES (2022) used the SVM model to predict MBTI based on Twitter posts which had a 0.84 accuracy. It indicates the limitations of this study. In future studies, it will be necessary to investigate more parameter settings to improve the performance of the classifier model.

The result of prediction indicates that males are more likely to be associated with "Thinking", while females have a more balanced distribution between "Thinking" and "Feeling", with a slightly higher proportion leaning towards "Thinking". Our findings do not exactly match those found in previous studies and are against the common stereotype. For example, Pavco Giaccia et al. (2019) found that participants were 1.3 times as likely to categorize a character as relating to rationality if it was preceded by a male as compared to a female prime. Sladek et al. (2010) reported that men preferred rational reasoning more than women, and conversely, women preferred experiential reasoning more than men.

In conclusion, the prediction results are against the common gender rationality stereotype and there is still space to improve the performance of our model. Has the evolution of society brought about changes in the perspectives of both men and women? If so, what factors are responsible for this shift? Are potential discrepancies in previous research methodologies the cause of any misinterpretations? Alternatively, could the limited data in this article influence the outcomes? Further research is needed to explore these questions and their underlying causes.

You are estimating the regression equation yi = $\beta 1 + \beta 2 * Di + \gamma * Xi + *(Xi - \hat{X})where Xi is a binary variable. Your estimates for \beta 2 and \delta is \beta 2 = 0.9 and \delta = 0.4 respectively. Calculate the estimate for the ATE 0 and X \hat{=} 0.2$

Acknowledgements

Iris Shi contributed to writing all sections of the paper and all coding work (data preprocessing, EDA

Model	Precision(T)	Recall(T)	F1(T)	Precision(F)	Recall(F)	F1(F)	Accuracy
RandomForest	0.74	0.95	0.83	0.79	0.36	0.50	0.75
SGDClassifier	0.83	0.93	0.88	0.82	0.62	0.70	0.82
LogisticRegression	0.82	0.92	0.87	0.79	0.61	0.69	0.81
LSTM	0.81	0.50	0.62	0.45	0.77	0.57	0.60

Table 3: Performance evaluation metrics of different models.

	Accuracy(T)	Accuracy(F)
Baseline	0.66	0.34

Table 4: Baseline of models.

plot making and model training).

Jikun Shen contributed to "1 Introduction", "3 Data" sections, spelling and grammar checking for the whole paper, EDA plot making, the coding of data preprocessing, organization and modification of all codes, and README file writing.

References

Isabel Briggs-Myers and Peter B. Myers. 1995. *Gifts differing: Understanding personality type*.

Steven Heine, Darrin Lehman, Kaiping Peng, and Joe Greenholtz. 2002. What's wrong with cross-cultural comparisons of subjective likert scales?: The reference-group effect. *Journal of personality and social psychology*, 82:903–18.

Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. pages 1706–1716.

Karen Jones. 2004. Gender and rationality. In Alfred R. Mele and Piers Rawling, editors, *The Oxford Handbook of Rationality*. Oxford University Press.

Ludovica Marinucci, Claudia Mazzuca, and Aldo Gangemi. 2023. Exposing implicit biases and stereotypes in human and artificial intelligence: State of the art and challenges with a focus on gender. *AI and Society*, 38(2):747–761.

CLEBER MARQUES. 2022. Kaggle.

MADHAV MATHUR. 2021. Kaggle.

Olivia Pavco Giaccia, Martha Fitch Little, Jason Stanley, and Yarrow Dunham. 2019. Rationality is Gendered. *Collabra: Psychology*, 5(1):54.

Barbara Plank and Dirk Hovy. 2015. Personality traits on Twitter—or—How to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, Lisboa, Portugal. Association for Computational Linguistics.

Ruth M. Sladek, Malcolm J. Bond, and Paddy A. Phillips. 2010. Age and gender differences in preferences for rational and experiential thinking. *Personality and Individual Differences*, 49(8):907–911.

Sanja Stajner and Seren Yenikent. 2021a. How to obtain reliable labels for MBTI classification from texts? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1360–1368, Held Online. INCOMA Ltd.

Sanja Stajner and Seren Yenikent. 2021b. Why is MBTI personality detection from texts a difficult task? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3580–3589, Online. Association for Computational Linguistics.

Bing Sun, Hongying Mao, and Chengshun Yin. 2020. Male and female users' differences in online technology community based on text mining. *Frontiers in Psychology*, 11.

RAJSHREE VARMA. 2021. Kaggle.