CLASSIFYING PERSONALITY TYPES

WITH

NEURAL NETWORKS

BASED ON

SOCIAL MEDIA (TWITTER) POSTS

AUTHOR: STEPHAN SCHULLER

ADVISED BY: DR. RAZVAN ANDONIE

(CWU) CS457

COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING

PROJECT OVERVIEW AND GOALS

- Make a Machine Learning model that uses a Neural Network
- Train the model using a Dataset of Status (tweets) that are paired with O.C.E.A.N. data
- Test the model on unpaired tweets to evaluate personality and credibility

INITIAL IDEA

- A bot, that can take a username
- Analyze with model
- Output some generalization

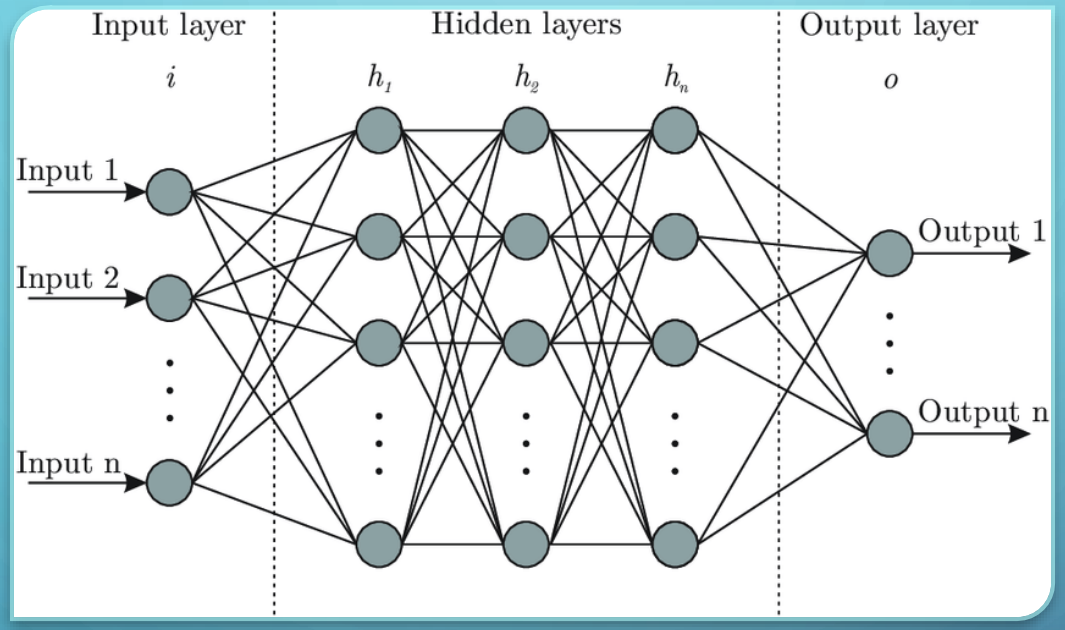
The personality type display here: INTP
Related to emotional volatility and patterns of dealing with stress poorly.
Highly neurotic people are more likely to experience negative emotions.

INSPIRATION

- Andrew Dunn
- Sentiment Analysis of Tweets
- Central Washington University - Ellensburg, Washington, United States
- CS 557

- Hima Vijay & Neenu Sebastian
- Personality Prediction using Machine Learning
- SCMS School of Engineering & Technology - Vidya Nagar, Karukutty, Ernakulam
- 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)

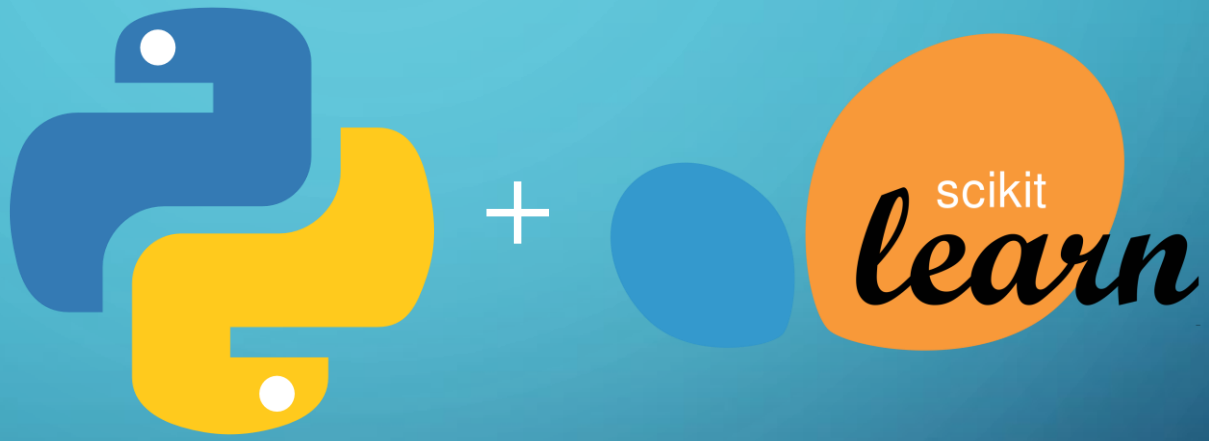
NEURAL NETWORKS



Not Natural Language Processing

TOOLS

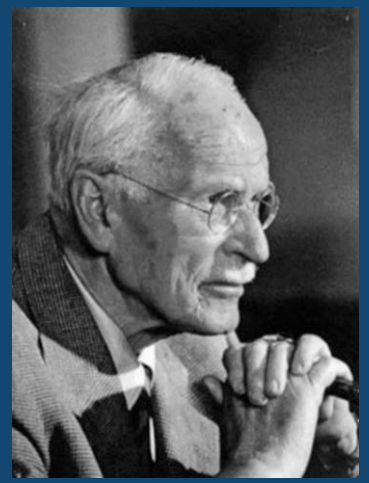
- Python 3.11
- SciKitLearn
- Pickle
- NLTK
- Tweepy



CLASSIFYING PERSONALITY TYPES O.C.E.A.N.

HISTORY OF PERSONALITY TRAITS

Carl Jung



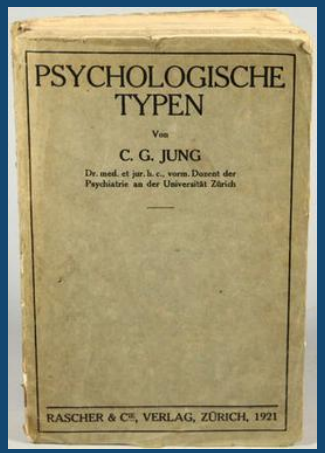
1875-1961

OBJECTIVE

1920's - 4000 measures of personality
1950's – 171 characteristics
1960's – 5 O.C.E.A.N.
1987's – 16 (based on 4) Myer's Briggs

SUBJECTIVE

Personality Types



1921

DIFFERENT VERSIONS OF THE SAME THING

Ocean		IPIP
Openness	>	Intellect
Conscientiousness	>	Conscientiousness
Extroversion	>	Extroversion
Agreeableness	>	Agreeableness
Neuroticism	>	Emotionality

HEXACO

Big 5, + H (H for Honesty / Humility)

Extraversion	X
Neuroticism	emotionality

Myers - Briggs

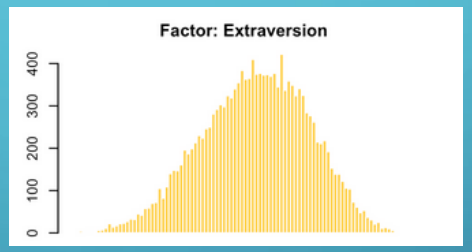
If Neuroticism is dropped
4 measures 2 polar
2^4 = 16

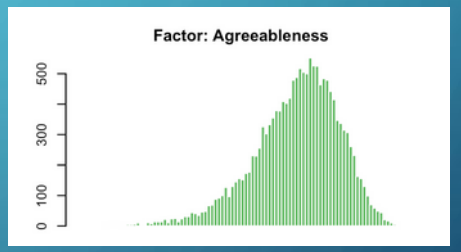
ESTJ, ENTJ, ESFJ, ENFJ,
ISTJ, ISFJ, INTJ, INFJ,
ESTP, ESFP, ENTP, ENFP,
ISTP, ISFP, INTP & INFP

The Datasets

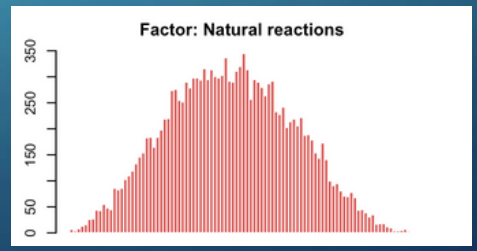
Open Psychometrics
IPIP FFM Dataset
1,015,342

My Personality Project
Stanford University
50,605





Personality in 100,000 Words
Tal Yarkoni
694



The Datasets (more details)

I am the life of the party.

- Disagree
- Slightly disagree
- Neutral
- Slightly agree
- Agree

```
EXT1    I am the life of the party.
EXT2    I don't talk a lot.
EXT3    I feel comfortable around people.
EXT4    I keep in the background.
EXT5    I start conversations.
```

All “I” Statements

[‘I’, ‘am’ , ‘the’ , ‘life’ , ‘of’ , ‘the’, ‘party’]

[‘life’ , ‘party’]

Words are sufficient to categorize

HOW CAN WE BE SO SURE?

Trait	No. of cats. (P < .05)	Top 20 LIWC categories	No. of words (p <.001)	Top 20 words
Neuroticism				
Anxiety	15	Feeling (0.17), Anxiety (0.16), Articles (-0.16), Space (-0.15), 1st Person Sing. (0.15), Certainty (0.13), 1st Person (0.12), Negative Emotions (0.12), Up (-0.11), Discrepancy (0.1), 2nd Person (-0.1), Affect (0.1), Negation (0.1), Grooming (0.1), Cognitive Processes (0.1)	33	awful (0.29), sick (0.26), road (-0.26), ground (-0.25), terribly (0.25), cranky (0.25), stress (0.24), feeling (0.24), southern (-0.24), stressful (0.24), myself (0.23), though (0.23), feel (0.23), sweater (0.23), county (-0.23), scenario (0.23), ashamed (0.22), feels (0.22), oldest (-0.22), spoiled (0.22)

Tal Yarkoni, *Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers*

DATA PREPROCESSING

Things that lowers accuracy of the model

- Misspelled words
- Hashtags
- URLs
- Uppercase / Lowercase
- Stopwords
- Unknown Characters
- Numbers
- Unique “words”

REDUCE NOISE to INCREASE ACCURACY

NLTK

“BAG-OF-WORDS” TACTIC 

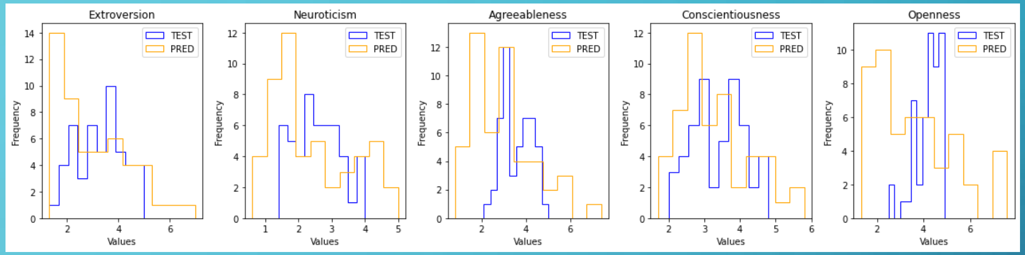
- Simplistic model that counts the occurrence of known words in a text
- My model uses a dictionary of 3,500 words
- A binary vector is constructed for each input text
  - Same Size as dictionary (each index is a word)
  - A value of 1 means the word exists

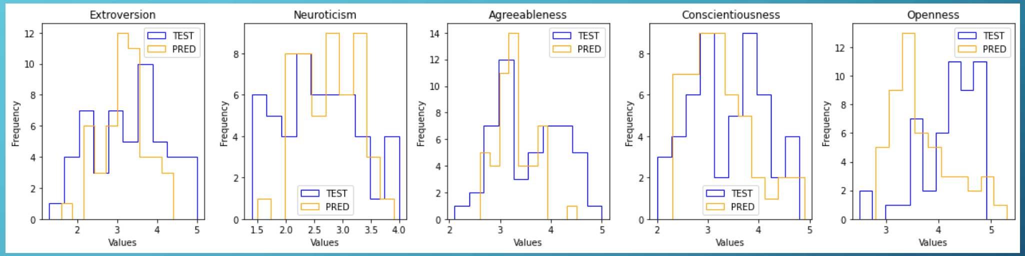
THE NEURAL NETWORK MODEL

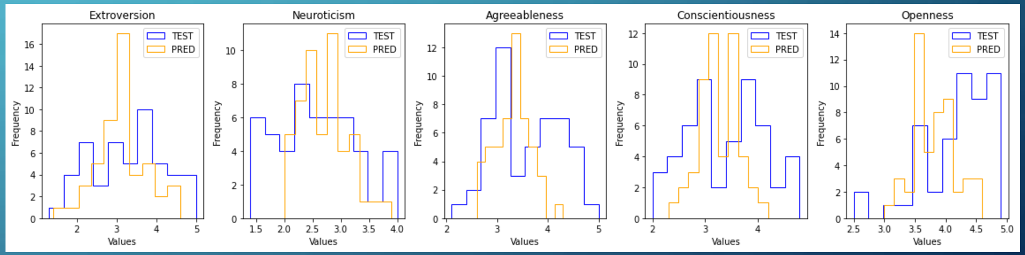
- Model is a standard Multi-Layer Perceptron
  - Mutli-Linear Output
  - 3500 input neurons
  - Two hidden layers, with 250 and 10 neurons
  - Uses relu function
  - Output layer is 5 neurons, and quantifies O.C.E.A.N.

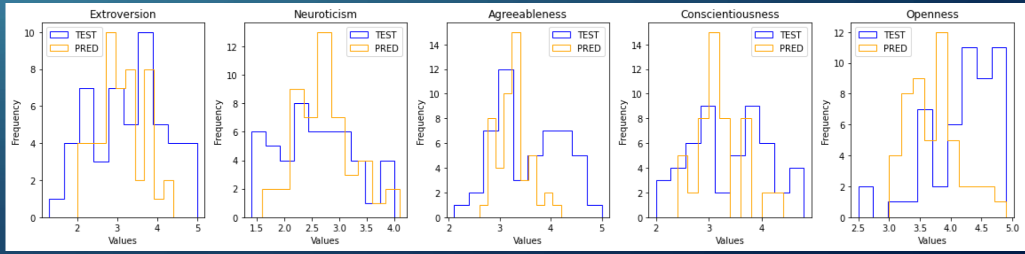
SOME RESULTS

TRIM SIZE	ITERATIONS	RMSE	HL1	HL2	SOLVER	ACT.
10,000	10,000	1.75565	100	10	adam	relu
5,000	5,000	1.65574	100	10	adam	relu
2,500	2,500	1.65412	100	10	adam	relu
2,000	1,000	1.41489	100	10	adam	relu
1.750	500	1.44395	100	10	adam	relu
1.750	500	DOES NOT CONVERGE	100	10	adam	relu
1,750	500	1.6005	100	10	adam	relu
1,750	500	DOES NOT CONVERGE	100	10	adam	relu
1,750	750	1.5766	100	10	adam	relu
1,500	750	1.6169	100	10	adam	relu
1,500	750	1.5396	100	10	adam	relu
1,500	750	1.5051	100	10	adam	relu
1,500	1000	1.4434	100	10	adam	relu
1,500	1000	1.4258	100	10	adam	relu
1,500	1000	1.5114	100	10	adam	relu
2,000	1000	1.4789	100	10	adam	relu
2,000	1000	1.4521	100	10	adam	relu
TRIM SIZE	ITER	RMSE	HL1	HL2	SOLV	ACT
2,000	1000	DOES NOT CONVERGE	100	10	lbfgs	relu
2,000	1000	DOES NOT CONVERGE	100	10	lbfgs	relu
2,000	2000	DOES NOT CONVERGE	100	10	lbfgs	relu
2,000	5000	1.0221	100	10	lbfgs	relu
2,500	5000	1.0159	100	10	lbfgs	relu
3,000	5000	0.9511	100	10	lbfgs	relu
5,000	5,000	0.7320	100	10	lbfgs	relu
5,000	5,000	0.8866	100	10	lbfgs	relu
5,000	5,000	0.9127	100	10	lbfgs	relu
3,000	5,000	1.0476	100	10	lbfgs	relu
4,000	4,000	0.8987	100	10	lbfgs	relu
3,500	4,000	0.9335	100	10	lbfgs	relu
3,500	4,000	1.0112	200	10	lbfgs	relu
3,500	4,000	1.489	300	10	lbfgs	relu
3,500	4,000	0.9439	234	10	lbfgs	relu
3,500	4,000	0.8720	234	10	lbfgs	relu
3,500	4,000	0.9761	234	10	lbfgs	relu

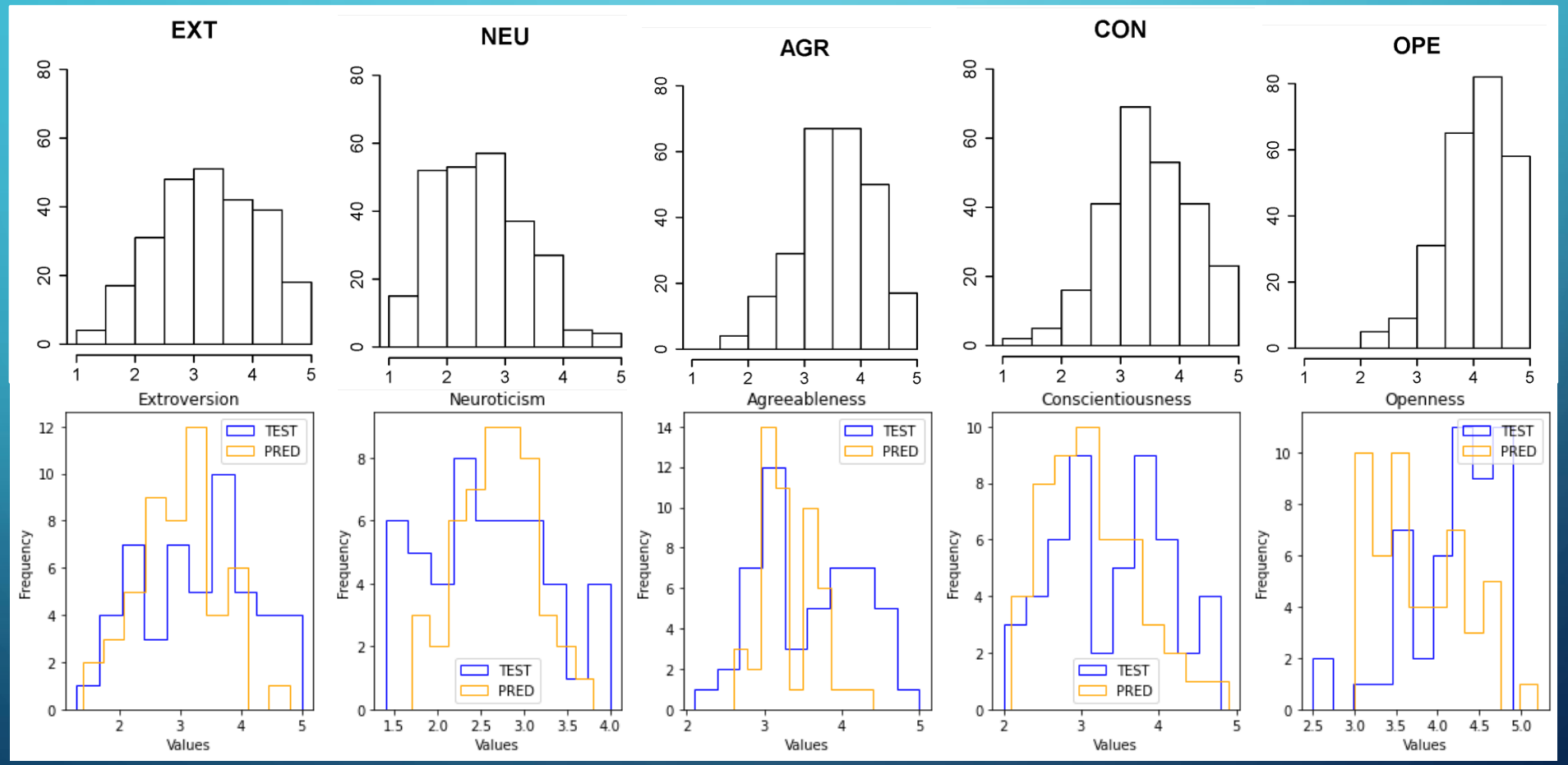








HOW CLOSE ARE WE?



IPIP

My Personality Project

Model

DISCUSSION

- Overall testing accuracy: 69%
- Bag-of-words tactic without NLP is has limiting results
- Perhaps wrong tools for the job
- NLP models have much higher accuracy (as high as 90%)
- Tweepy API is expensive (now)
- Need a fairly large sample
- Accurate evaluation needs 300 to 500 words per Author for accuracy above 70%

REFERENCES

- Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26
- Hima Vijay & Neenu Sebastian (2022) *Personality Prediction using Machine Learning* - 2022 International Conference on Computing, Communication, Security and Intelligent Systems https://searchlib.cwu.edu/permalink/01ALLIANCE_CWU/1c5n89p/cdi_globaltitleindex_catalog_367322427.
- Andrew Dunn 2020, *Sentiment Analysis of Tweets* – Central Washington University
- Tal Yarkoni, 2010, *Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers,* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2885844/