# Demographic Inference TASKS

- ☑ Preprocessing with profile info
  - ☑ Clean json and prepare csv
  - ☑ Run Roberta Model
- ☑ Compare new DA prediction w/ new ones
- ☑ Drug Type Identification
  - ☑ Add multiple drug types if exists
- ☑ Demographic Inference
  - ☑ Demo inference script
- ○ Span Emo
  - ☑ Validate demo inference Span Emo
  - ☑ tweak script for custom CSV
  - ☑ Span Emo validation
    - ○ F1-Micro score: 70%
  - ○ Span Emo check/join for very large data
  - ☑ Script to do it for multiple folders at once
  - ○ Data from https://competitions.codalab.org/competitions/17751#learn_the_details-datasets
- ○ VADER
  - ☑ Single script that predicts sentiment
  - ☑ Add sentiment score in vader script
  - ○ VADER validation
    - ☑ VADER (F1 = 0.96) actually outperforms even individual human raters (F1 = 0.84) at correctly classifying the sentiment of tweets.
    - ○ sentiment['compound'] < -0.05 => Negative.
      sentiment['compound'] > 0.35 => Positive.
- ○ RACE and Ethnicity
  - ○ https://github.com/appeler/ethnicolr
  - ○ https://github.com/wri/demographic-identifier also uses ethnical
- ○ BERT topic analysis
  - ☑ Train; get topic

  - ○ GPT prompt; how?
  - ○ https://www.pinecone.io/learn/bertopic/
  - ○ https://maartengr.github.io/BERTopic/api/representation/openai.html
  - ○ **https://github.com/MaartenGr/BERTopic**
- ○ Multivariate analysis
  - ○ Methods study??
- ○ Manuscript
  - ✓ Skeleton
  - ○ Abstract
  - ○ Literature Review
  - ○ Methods
  - ✓ Data Preprocessing
  - ○ Data Flow Diagram
    - ○ References
    - ○ Social Media Drug Abuse and age and gender
    - ○ Social media and Age and Gender
    - ○ Emotional tone, sentiment, theme/topic
    - ○ Large scale social media profile
  - ○ Quantitative analysis

| Year | Preprocess w/ profile info | Re run Drug abuse prediction | Demo infer | Sentiment infer | Emo infer | Topic Analysis | Dump in single file | Quantitative | Quantitative |
|---|---|---|---|---|---|---|---|---|---|
| 2019 | done | done | done | Done | Doing | | done | | |
| 2020 | done | done | done | done | done | | done w/ sentiment | done | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2021 | done | done | Done | done | doing | | Done | | |

| ***2019_old*** | ***2019_new*** | ***2020_old*** | ***2020_new*** | ***2021_old*** | ***2021_new*** |
|---|---|---|---|---|---|
| 01 --> (244486, 3) | 01 --> (314802, 13) | 01 --> (298034, 3) | 01 --> (298598, 14) | 01 --> (102294, 3) | 01 --> (102480, 13) |
| 02 --> (261101, 3) | 02 --> (262253, 13) | 02 --> (133860, 3) | 02 --> (133950, 14) | 02 --> (257633, 3) | 02 --> (258027, 13) |
| 03 --> (230673, 3) | 03 --> (243710, 13) | 03 --> (348130, 3) | 03 --> (340683, 14) | 03 --> (264017, 3) | 03 --> (264285, 13) |
| 04 --> (267526, 3) | 04 --> (264268, 13) | 04 --> (349761, 3) | 04 --> (351455, 14) | 04 --> (115620, 3) | 04 --> (115718, 13) |
| 05 --> (283141, 3) | 05 --> (284107, 13) | 05 --> (300394, 3) | 05 --> (302246, 14) | 05 --> (233142, 3) | 05 --> (233402, 13) |
| 06 --> (279331, 3) | 06 --> (287253, 13) | 06 --> (330517, 3) | 06 --> (331107, 14) | 06 --> (224678, 3) | 06 --> (239514, 13) |
| 07 --> (222454, 3) | 07 --> (223417, 13) | 07 --> (277008, 3) | 07 --> (277488, 14) | 07 --> (223018, 3) | 07 --> (223202, 13) |
| 08 --> (199560, 3) | 08 --> (200202, 13) | 08 --> (335009, 3) | 08 --> (335339, 14) | 08 --> (238446, 3) | 08 --> (238584, 13) |
| 09 --> (194107, 3) | 09 --> (194789, 13) | 09 --> (307877, 3) | 09 --> (308060, 14) | 09 --> (218731, 3) | 09 --> (218902, 13) |
| 10 --> (176767, 3) | 10 --> (177181, 13) | 10 --> (329109, 3) | 10 --> (329913, 14) | 10 --> (238427, 3) | 10 --> (263079, 13) |
| 11 --> (188509, 3) | 11 --> (188957, 13) | 11 --> (245544, 3) | 11 --> (245858, 14) | 11 --> (217406, 4) | 11 --> (224491, 13) |

| | | | | | |
|---|---|---|---|---|---|
| 12 --> (306368, 3) | 12 --> (307397, 13) | 12 --> (263789, 3) | 12 --> (264259, 14) | 12 --> (234558, 3) | 12 --> (235780, 13) |
| **(2854023, 3)** | **2948336, 13** | **(3519032, 3)** | **(3518956, 15)** | **2567970, 3** | **2617464, 13** |
| | 2799726, 16 | | 3502171, 16 | | 2553235, 16 |

Old total: (8941025, 4)