Personality Prediction from Twitter Data

Agenda

- Motivation
- Data Collection
- Data Pre processing (Filtering)
- Feature Engineering
- Modeling - Supervised and Unsupervised
- Geo Visualization
- Comparison of Modeling Approaches
- Conclusion

Motivation

- Personality trait can be defined as habitual patterns of behavior, thought & emotion
- People express views on social media (tweets, blogs etc.)
- Social activity can be used to deduce personality traits
- “Processing User Tweets to identify Personality Types”
- Considerations
  - 10 Personality categories [Personality_Traits]
  - Geo tagged data for visualization [Geo_Plots]

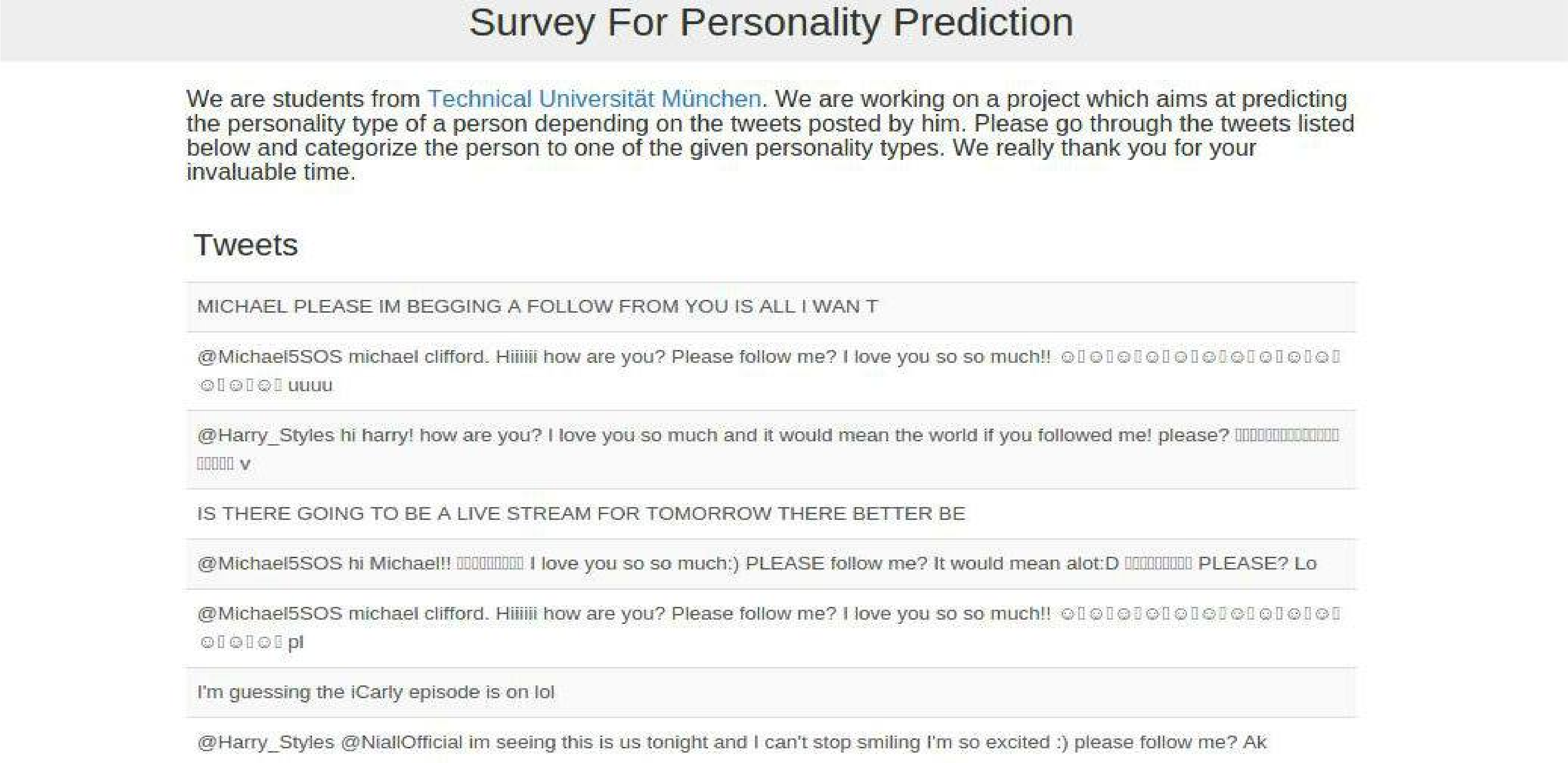
Personality Types

- Conscientiousness
- Extrovert
- Agreeable
- Empathetic
- Novelty Seeking
- Perfectionist
- Rigid
- Impulsive
- Psychopath
- Obsessive

Data Collection

- Data not labeled according to personality type.
- Possible Solutions
  - Manually process data to assign labels (Time!)
  - Survey Form (No platform for dynamic data)
- Survey Application for labeling the data
  http://survey.lostarray.xyz

Survey Application



Survey For Personality Prediction

We are students from Technical Universität München. We are working on a project which aims at predicting the personality type of a person depending on the tweets posted by him. Please go through the tweets listed below and categorize the person to one of the given personality types. We really thank you for your invaluable time.

Tweets

MICHAEL PLEASE IM BEGGING A FOLLOW FROM YOU IS ALL I WAN T

@Michael5SOS michael clifford. Hiiiiii how are you? Please follow me? I love you so so much!! ☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯ uuuu

@Harry_Styles hi harry! how are you? I love you so much and it would mean the world if you followed me! please? ▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯ v

IS THERE GOING TO BE A LIVE STREAM FOR TOMORROW THERE BETTER BE

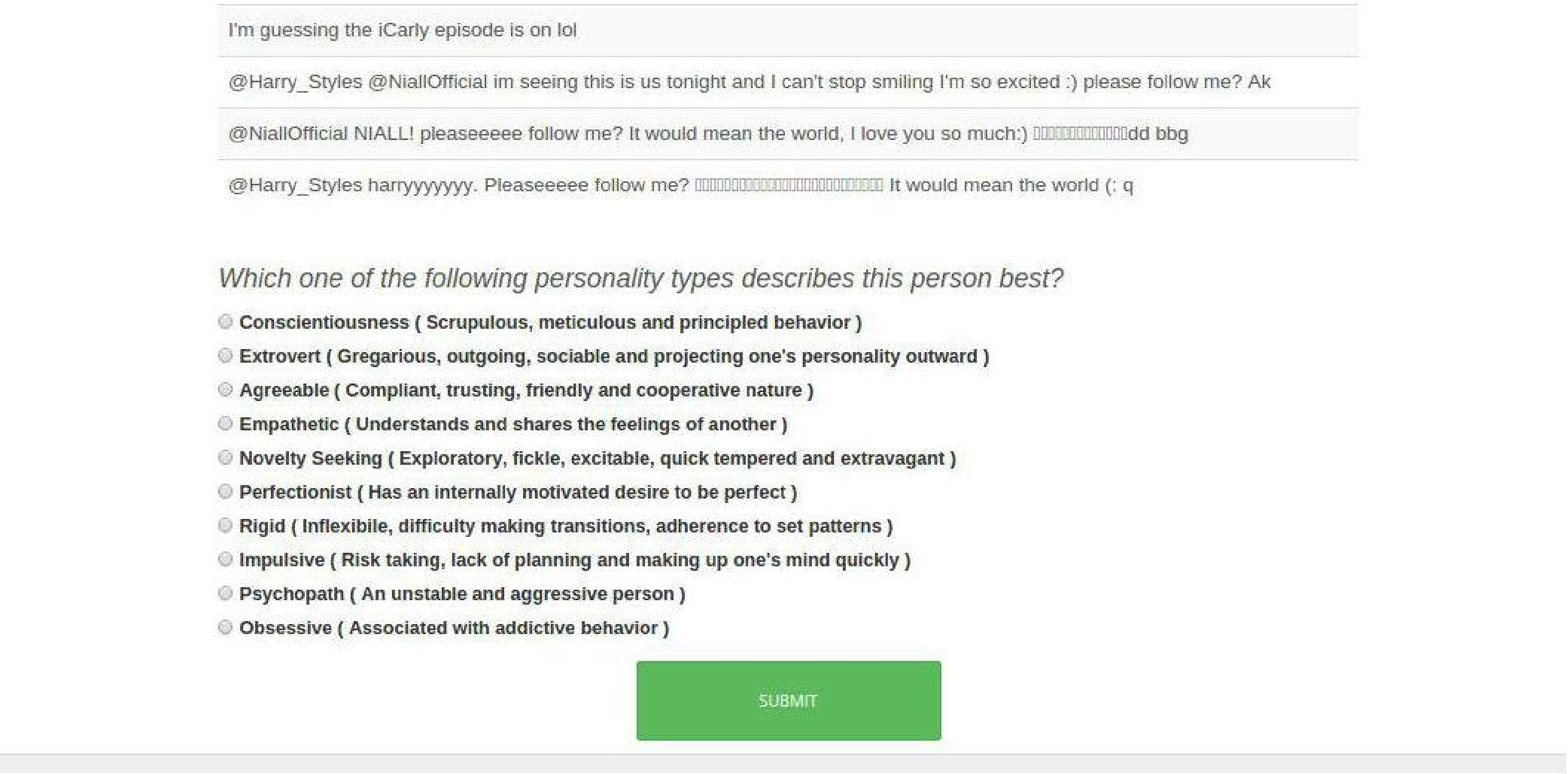
@Michael5SOS hi Michael!! ▯▯▯▯▯▯▯▯▯ I love you so so much:) PLEASE follow me? It would mean alot:D ▯▯▯▯▯▯▯▯▯ PLEASE? Lo

@Michael5SOS michael clifford. Hiiiiii how are you? Please follow me? I love you so so much!! ☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯☺▯ pl

I'm guessing the iCarly episode is on lol

@Harry_Styles @NiallOfficial im seeing this is us tonight and I can't stop smiling I'm so excited :) please follow me? Ak

Survey Application (2)



I'm guessing the iCarly episode is on lol

@Harry_Styles @NiallOfficial im seeing this is us tonight and I can't stop smiling I'm so excited :) please follow me? Ak

@NiallOfficial NIALL! pleaseeeee follow me? It would mean the world, I love you so much:) □□□□□□□□□□□□□dd bbg

@Harry_Styles harryyyyyyy. Pleaseeeee follow me? □□□□□□□□□□□□□□□□□□□□□□□□□□ It would mean the world (: q

Which one of the following personality types describes this person best?

- ○ **Conscientiousness (Scrupulous, meticulous and principled behavior)**
- ○ **Extrovert (Gregarious, outgoing, sociable and projecting one's personality outward)**
- ○ **Agreeable (Compliant, trusting, friendly and cooperative nature)**
- ○ **Empathetic (Understands and shares the feelings of another)**
- ○ **Novelty Seeking (Exploratory, fickle, excitable, quick tempered and extravagant)**
- ○ **Perfectionist (Has an internally motivated desire to be perfect)**
- ○ **Rigid (Inflexibile, difficulty making transitions, adherence to set patterns)**
- ○ **Impulsive (Risk taking, lack of planning and making up one's mind quickly)**
- ○ **Psychopath (An unstable and aggressive person)**
- ○ **Obsessive (Associated with addictive behavior)**

Data Preprocessing

- Analysis of Tweets
  - SPAM

    e.g. ☀☀☀☀☀☀☀☀☀☀☀☀☀☀☀ Hi @zaynmalik Please follow me It would mean the world I can't wait to see This Is Us!! ☀☀☀☀☀☀☀☀☀☀☀☀☀☀☀ 1
  - Retweets

    e.g Princess Diana: British Police Investigating New Leads in Death @MJJPEACE
  - Genuine

    e.g. @EsperTortuga "oH GOD IT'S HORRIFYING"
- Filter Tweets
  - Based on frequency of tweets per user
  - Tweets with hyperlinks
- Remove Stopwords (and, the, a, etc.) [NLTK]

Feature Engineering

- Bag of words
- N grams (Shingles)
- Ensure full word coverage in training & evaluation dataset
- Term Frequency – Inverse Document Frequency (TF-IDF)
- One Hot Encoding (OHE)
- Sentiment [Text_Blob]
- Frequency of Tweets

Modeling - Supervised Learning

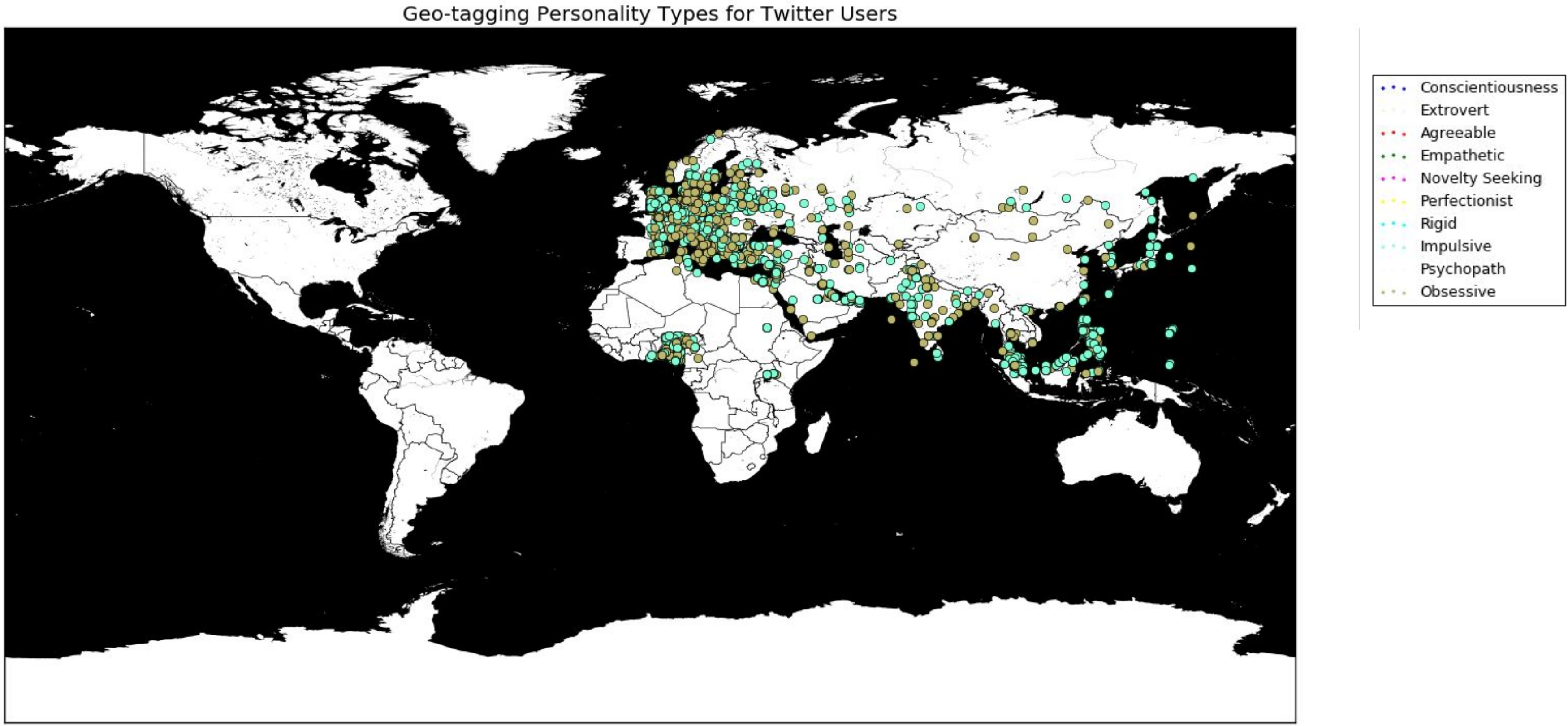
- Features
  - Sentiment, words, frequency of tweets
- Random Forest
- kNN
- Neural Networks
- Multiclass SVM
- Stratified k-fold cross validation

Supervised Model Performance

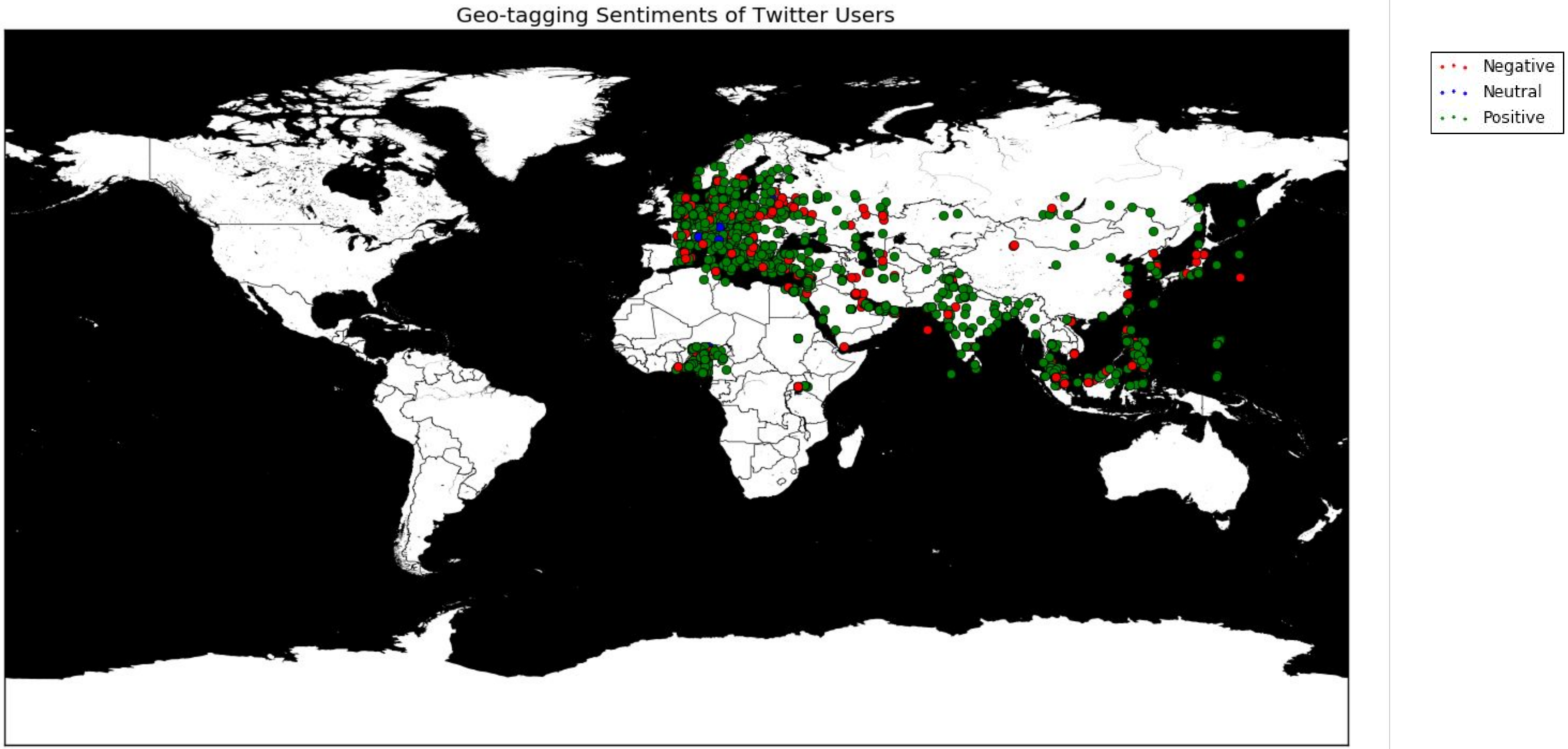
- Bag of words (~39000 features)
- NGrams (~600K features)
- Dimensionality Reduction
- PCA (Dim=5000)
- Label data is too less to evaluate any performance measure (~100)
- All models give same performance (Accuracy)

  Because of choice of features (quality of tweets)

Geo Visualization I - Personality Types



Geo Visualization II - Sentiment

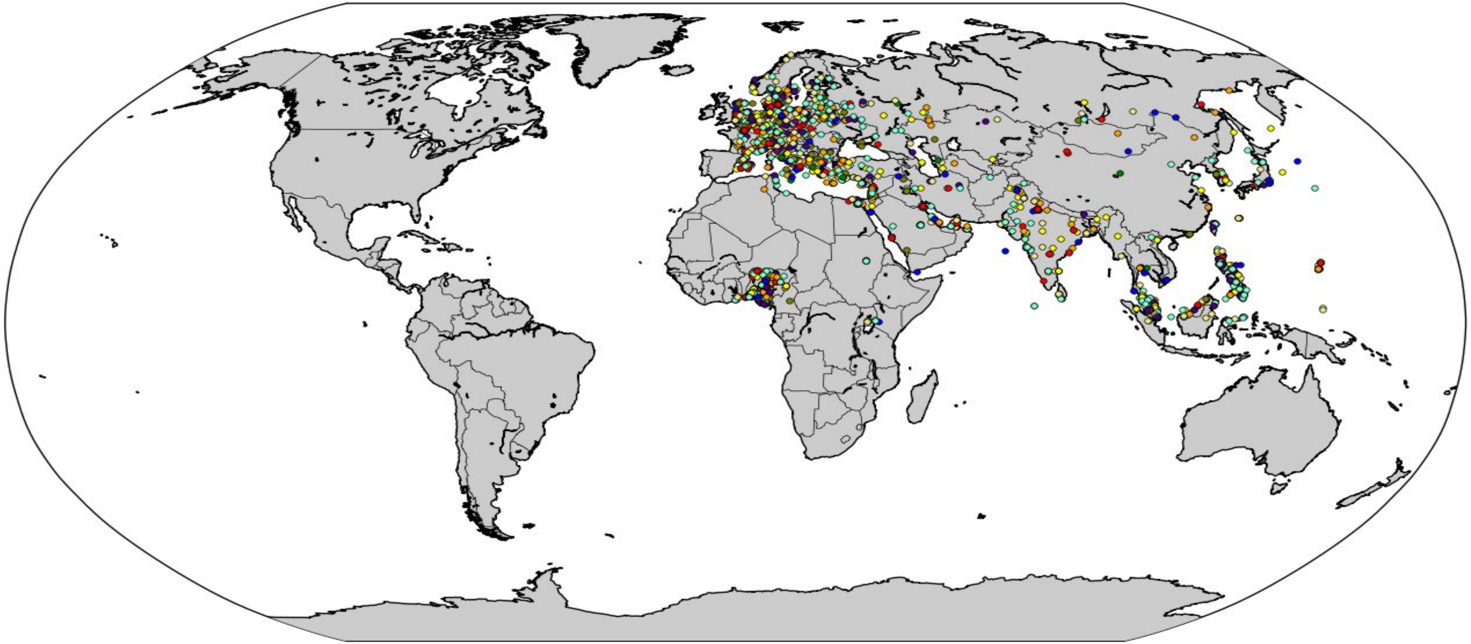


Geo-tagging Sentiments of Twitter Users

Unsupervised Model

- k-means clustering for 10 clusters
- Features used
    - Bag of words (adjectives)
    - Polarity of tweets
    - Subjectivity [Text_Blob]
    - Tweet frequency of a user
- Principal Component Analysis (dimensions =2000)

k-means Clustering



Comparison of Modeling Approaches

- Mapping clusters of Unsupervised models to Personality Types from Supervised model.
- Supervised
  - Prediction: Only 2-3 types
  - 50% of Training concentrated on these types (due to quality of tweets)
- Unsupervised
  - Prediction: 10 types (used same features)

Conclusion

- Problems & Challenges
  - Quality of data
  - Choice of features
- Applications & Use Cases
  - Market strategy for specific region
  - Advertisements - Decide time & target audience
  - Use same model to process user data from another social platform (e.g. Quora)
- Future Improvements
  - Increase Labeled data
  - NLP methods for advanced analysis to extract semantic features from tweets
  - Temporal analysis of tweets

Bibliography

- https://en.wikipedia.org/wiki/Trait_theory#List_of_personality_traits [Personality_Traits]
- http://www.nltk.org/ [NLTK]
- https://textblob.readthedocs.org/en/dev/ [Text_Blob]
- http://tweettracker.fulton.asu.edu/tda/TwitterDataAnalytics.pdf
- http://matplotlib.org/basemap/ [Geo_Plots]
- Code for Twitter User Personality Prediction
- Code for Survey Application

Questions?

THANK YOU!