Tweet, Reaction, Action

Brian van der Bijl

Outline

Introduction

What am I doing again?

- Investigating AI based preselection for incitement detection in newsrooms
- Starting point: apply *NLP* and *sentiment analysis*
- Follow-up: analyse structure and flow of conversations

Plan: How it Started

- Get dataset
- Train BERT
- Predict tweet-to-conversation mappings
- Profit!

Plan: How it's Going

- Get dataset
  - Collect tweets and conversations
  - Tag data (using TAs)
  - Analyse results
  - Yield usable dataset
- ~~Train BERT~~
- ~~Predict tweet-to-conversation mappings~~
- Profit?

What's New

Focus has shifted towards delivering a reusable and well-understood dataset for research on intent-informed incitement detection in Twitter data.

New Goals and Deliverables

What is Being Built

- Intial set containing 500 tweets (roughly) on Dutch COVID-19 response
- Second set containing (as of now) 356 tweets on Dutch response to Ukraine war
- Tweets form chains or more complex trees

What is Being Built

- Tweets tagged on 15 axes by 6 TAs (aiming for 100% coverage) and additional volunteers
  - 3 TAs from Media/Journalism
  - 3 TAs from ICT
- Average of tags for each tweet generates embedding in 15-D space

15-D Intent Space

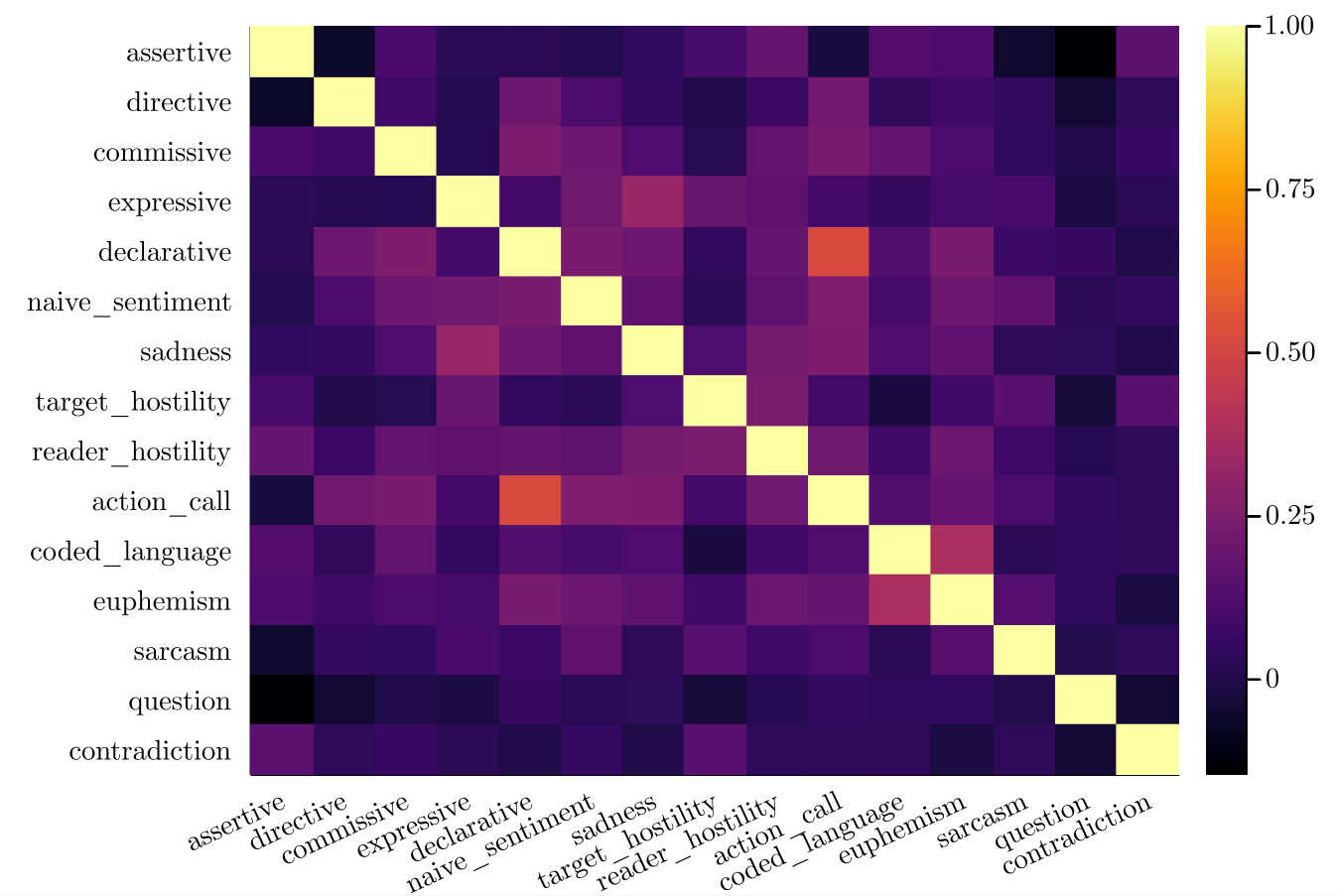
- Assertiveness
- Directiveness
- Commissiveness
- Expressiveness
- Declarativeness
- Naive sentiment
- Sadness
- Hostility towards target
- Hostility towards reader
- Call for action
- Use of coded language
- Use of euphemism
- Use of sarcasm
- Question
- Contradiction

Analysis

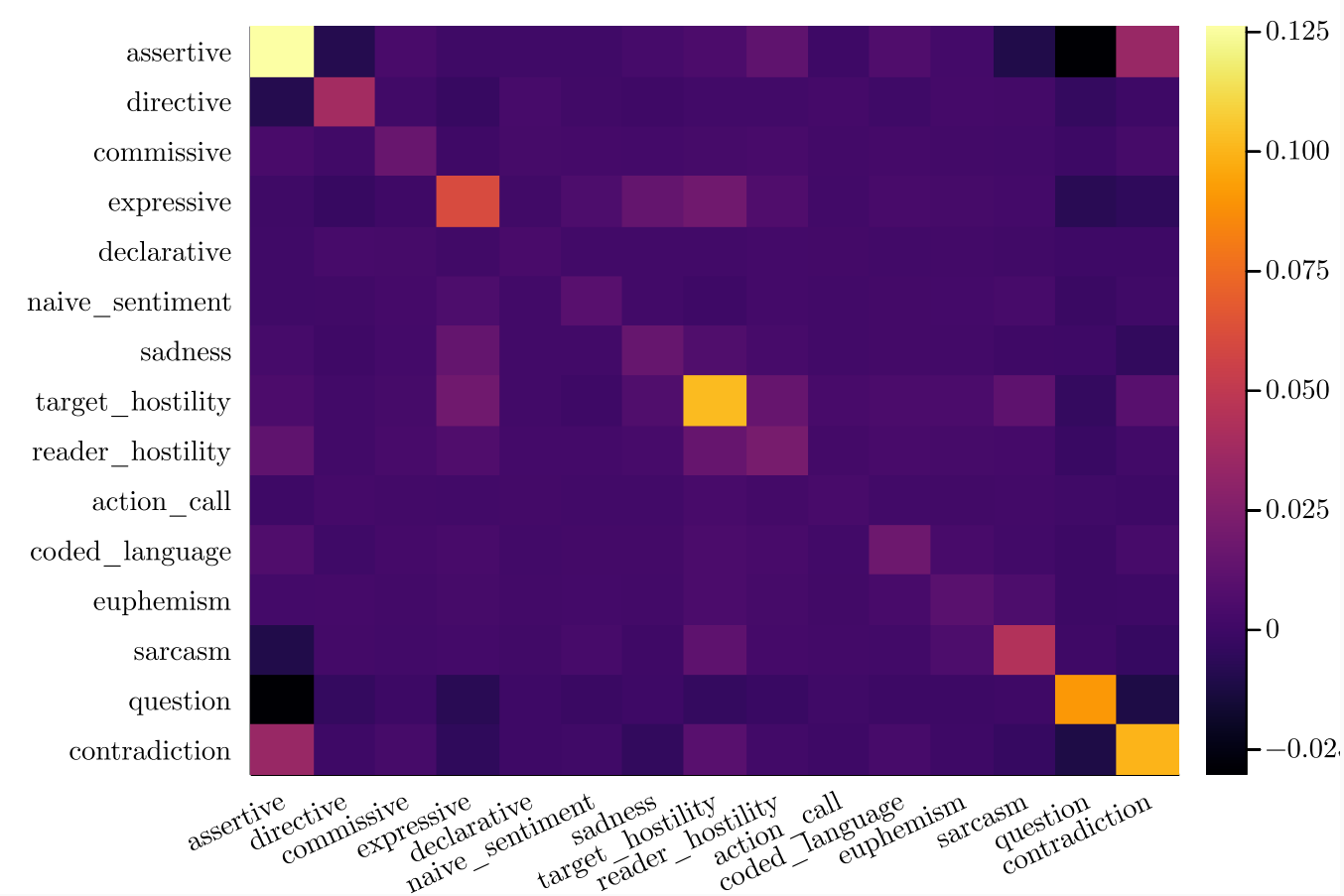
Analysis of Chosen Basis

- Current tagged embeddings use a space with 15 basis vectors
  - This might include redundancy
- Analyse correlation between different basis vectors
  - How many do we need to keep to capture enough data for ML algorithms?
  - How many do we need to maintain explainability?
- Can we easily map from 15-D explainable tags to lower dimensional approximations?

Correlation



Covariance


assertive
directive
commissive
expressive
declarative
naive_sentiment
sadness
target_hostility
reader_hostility
action_call
coded_language
euphemism
sarcasm
question
contradiction
0.125
0.100
0.075
0.050
0.025
0
−0.02

Space Density Analysis

- What areas of the space are uninhabited?
  - Are we missing certain classes of tweets?
  - Are certain combinations of values “impossible”?
- Does the data show clear separation between clusters?

How?

- Clustering in 15-D is challenging
- Application of Factor Analysis to detect smaller set of equivalent intuitive indicators
- Clustering and downscaling using t-SNE algorithm
- Gaussian Mixtures on lower dimensional results

Bias Analysis

- Distinction between to types of students
  - Do these groups tag differently?
  - In-group variance vs inter-group variance
  - Can we correct for these biases?
- What biases cannot be compensated within this set?
  - E.g. level of education

Current Status

Current and Projected Progress

- First tagger finished first 500 tweet set with hours to spare
- One tagger started just recently
- Rest is on track / projected to stay (well) within budget
- Tagging of second set projected to be within or slightly above budget

Recommendations

- Invest, if needed, some additional hours to get 100% coverage on second set
- Make both sets available as basis for future research on incitement detection

Future

- Expand set of taggers and aim for a more diverse background
- Training BERT or similar could be done as a (bachelor) student project
  - Automated data generation / augmentation
  - Crucial step in PoC deployment
- Conversation analysis as future research
  - Mapping embeddings to potential successor node embeddings
  - Semantic clusters as states in Markov chain

Questions?