Reddit Content Analysis

Group 22

Markus Ritzer	Data Preprocessing, Submission Clustering
Thomas Aumüller	Community and Throw-away Analysis
Oana-Emilia Bodirnea	Sentiment Analysis
Gregor Yinan Potthast	Sentiment Analysis

Motivation

Analyze recent activities in r/conspiracy subreddit

- Topic Clustering
  - Comparison of TFIDF and Word Embeddings
  - Presentation of Clusters as Word Clouds
- Community Analysis
  - Track activity of users
  - detect Throw-Away accounts
- Sentiment Analysis
  - Detect positive/negative/neutral sentiment of
    - submissions
    - comments

Data and Methods

Data Crawling

- PRAW Library
  - 1k latest/hottest submissions
  - unique authors
  - comments
- OpenAI
  - Word Embeddings
- Hugging Face
  - Sentiment analysis

Methods

- TFIDF
- Word embeddings
- K-Means clustering
- Mannwitneju significance test
- General preprocessing
  - spell correction
  - stopword removal
  - punctuation removal
  - lemmatization

Results: Topic Clustering

Submission to Vector via average pooling of

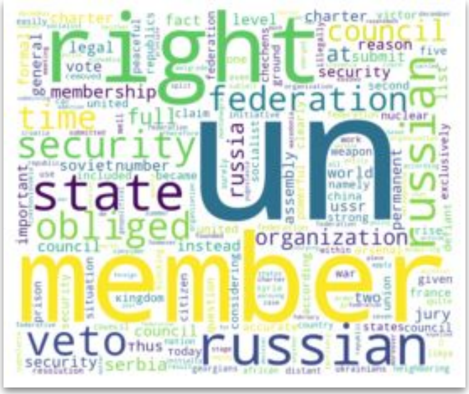
- TFIDFs (TfidfVectorizor)
- Word Embeddings (OpenAI text-embedding-ada-002 model)

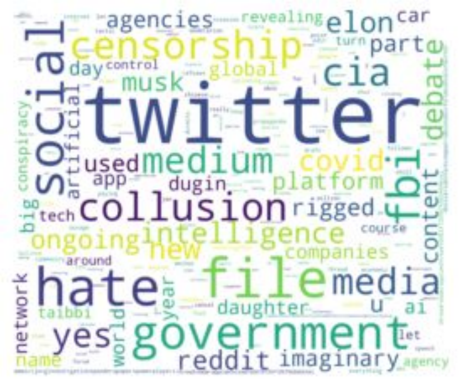
Clustering with K-Means (k = 30)

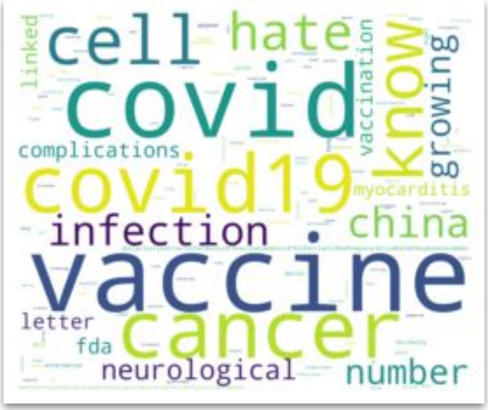
- Averaging pairwise cosine similarities within clusters
- Averaging over cluster similarities

TFIDFs: **8.36 %**

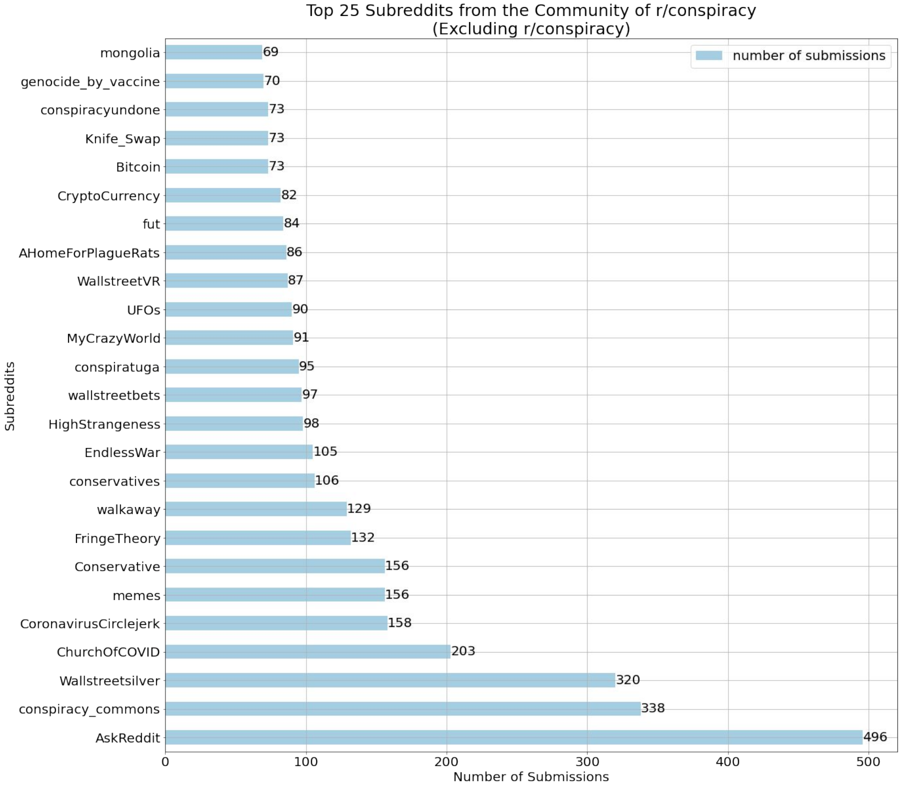
Word Embeddings: **97.10 %**







Results: Community Analysis / Throwaway Analysis



Throwaway Account(TA) Def.:

First Submission of the Redditor was in the observed Subreddit (r/conspiracy).

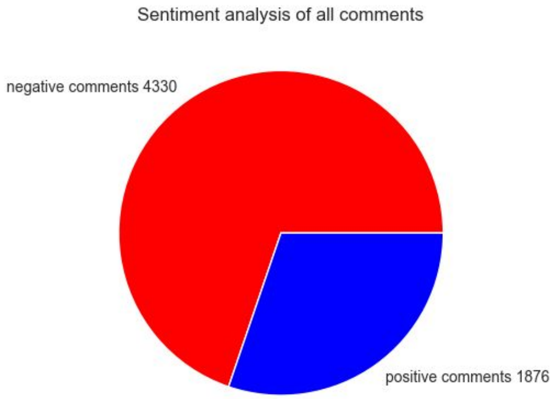
Results:

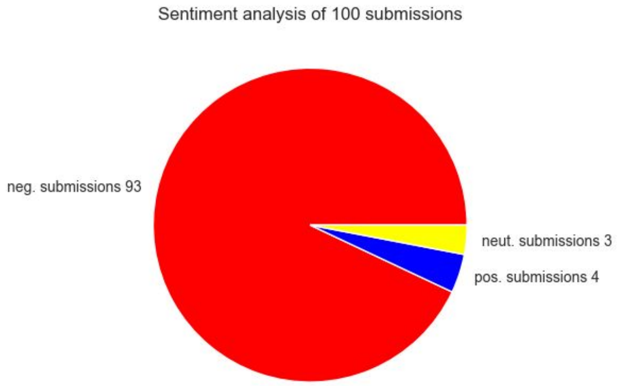
38.57% of our user base potentially used throwaway accounts.

There was no significant difference in user-activity when comparing normal users and TAs with MWU-Test.

Results: Sentiment Analysis

- Analysis of the comments from the top 100 submissions with the most comments





Conclusion

Conclusion

Data collection framework allows variety of analyses
Word Embeddings working better than TFIDF
Found activity behaviour and sentiment

Future work

adapt data collection framework for more precise retrieval

Limitations/Biases

- find best k for Clustering
- obtain adequate word embeddings
- more precise preprocessing (e.g. remove links)
- PRAW retrieval limits
- OpenAI retrieval limits