# Detection of Antisocial Behavior in Online Communities

Martin Borák, Supervisor: Ivan Srba

## Online communities

- social networks, knowledge sharing systems, online games, news and entertainment portals
- hundreds of millions people in the world
- user generated content
- antisocial behavior: haters, trolls, flamers, spammers, cyberbullies
- regulated mainly by moderators – goal is to make their job easier

## Detection of content containing hate in YouTube comment sections

### Data acquisition

- over 200 000 comments collected from political YouTube channel The Young Turks
- YouTube API + JavaScript

### Labeling of data

- over 6 000 comments chosen for labeling
- crowdsourcing via custom Django app
- 700 comments labeled as either hateful or benign by 24 participants
- weighted average Fleiss kappa of 0.6813

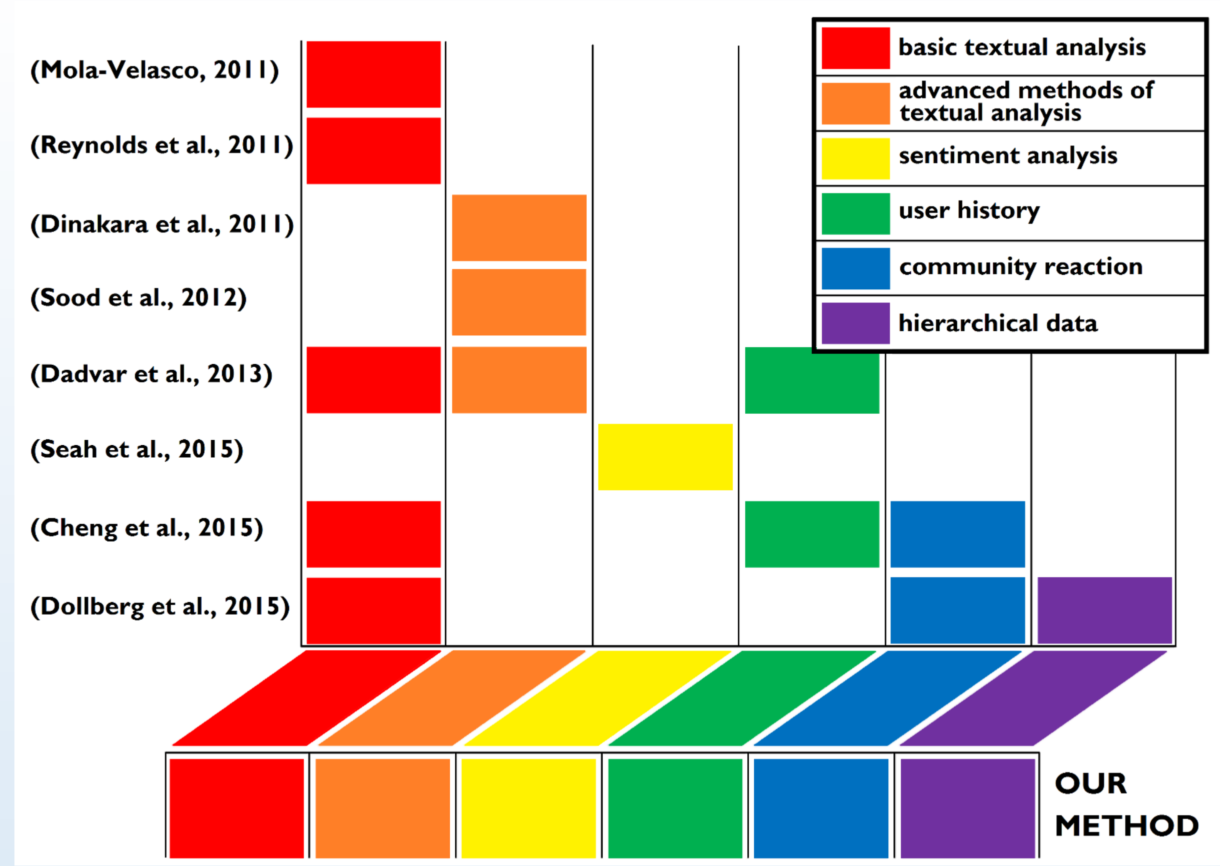
### Classification using machine learning

- data preprocessing – lemmatization, stop-words removal
- feature extraction (extracted 117 features, 64 of them used in classification)
- min-max normalization, oversampling
- different supervised classifiers
- k-fold validation
- parameter tuning
- problem – not enough labeled data
- co-training – semi-supervised machine learning method

STU FIIT

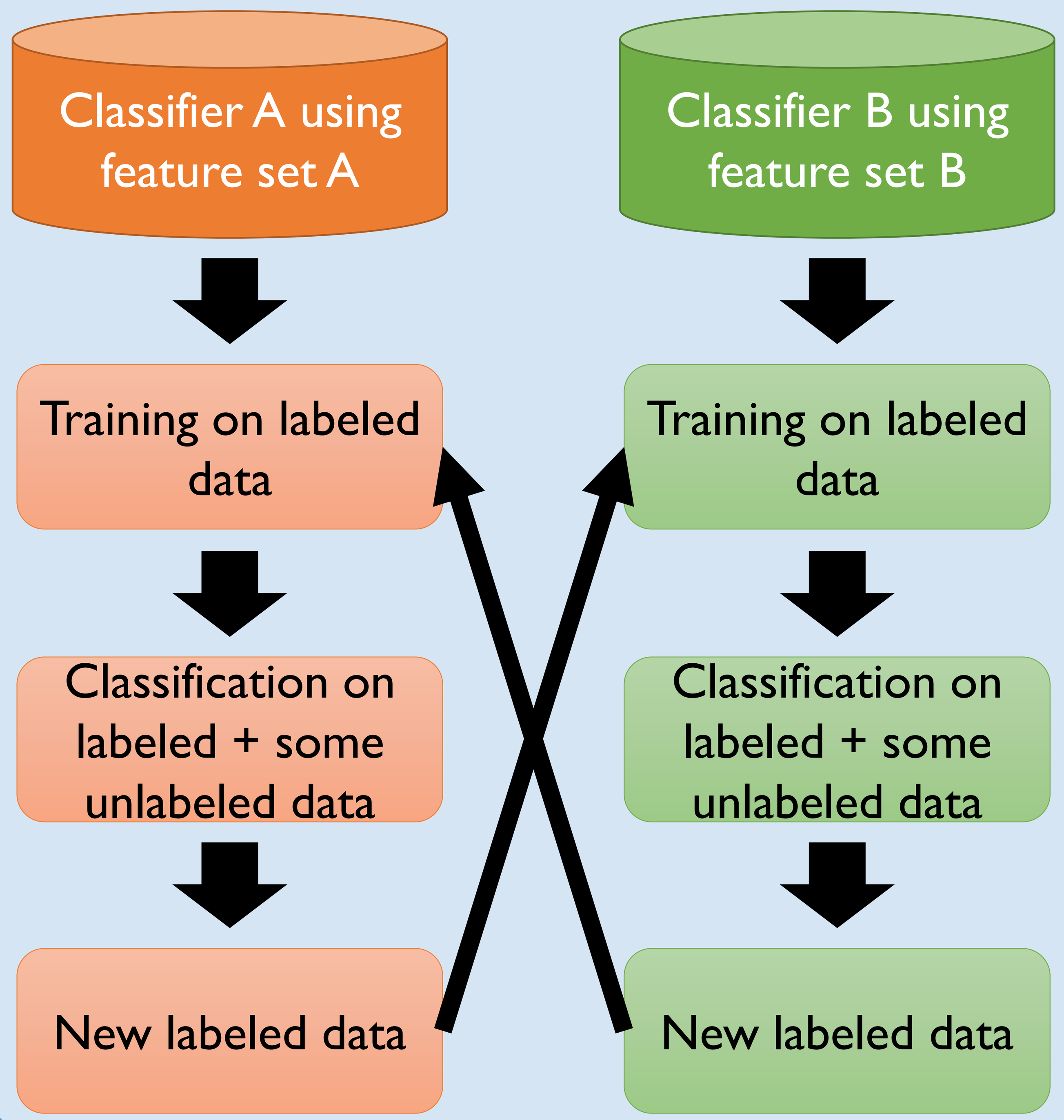
## Our method for automatic detection of antisocial behavior

- existing solutions differ in platforms, type of behavior that is being detected, utilized algorithms and used features
- detection of antisocial users vs. **detection of inappropriate content**
- machine learning based approach
- different categories of features:
  - textual features (including sentiment analysis)
  - features of user history
  - community reaction
  - hierarchical data
- **our hypothesis:** by combining features from all feature categories, the ability to detect antisocial behavior increases




## Co-training

- two classifiers use two different sets of features
- feature sets must be independent and uncorrelated
- co-training uses unlabeled data to construct new labeled data for future iterations



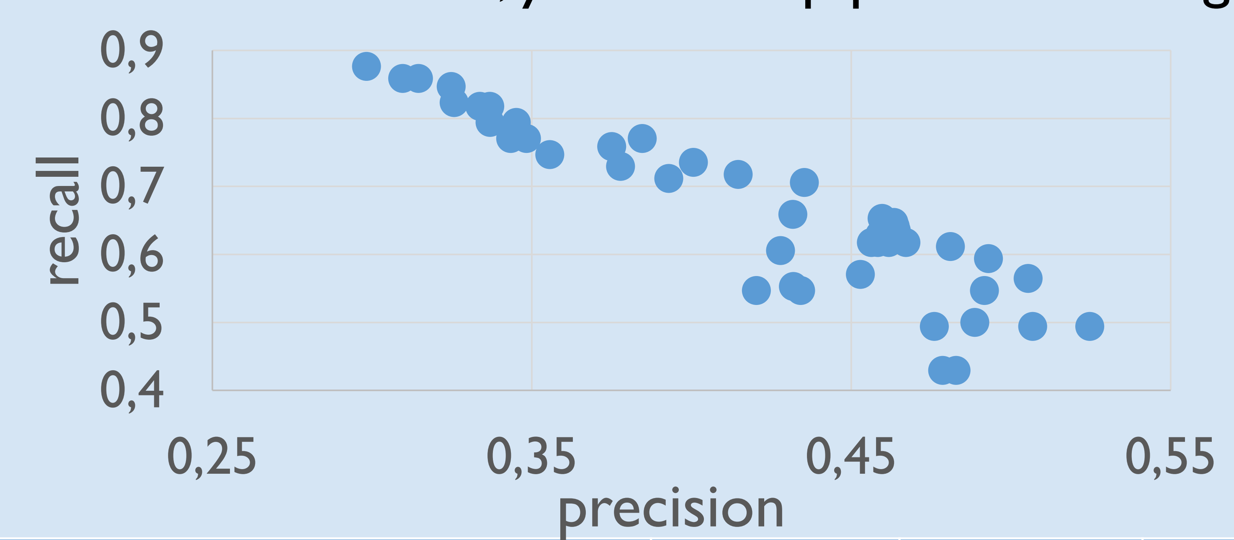

## Results

### Supervised classifiers

- best results with Extremely randomized trees classifier (ERT)
- classification using different combinations of feature categories:

### Co-training

- best combination of classifiers and feature sets:
  1. Extremely randomized trees (textual features + hierarchical data)
  2. AdaBoost (user history features + community reaction)
- results vary for different parameter settings of co-training algorithm
- our goal is to maximize recall, yet still keep precision as high as possible




| Type of classification | Precision | Recall | $F_1$-score |
|---|---|---|---|
| ERT – textual features + hierarchical data | 39.24 % | 55.53 % | 45.13 % |
| ERT – user history + community reaction | 39.30 % | 57.80 % | 46.02 % |
| ERT – all | 45.76 % | 58.00 % | 50.00 % |
| co-training (with highest $F_1$-score) | 46.34 % | 64.71 % | 53.56 % |

## Conclusion

- results confirm, that combination of all feature categories trains a classifier better then a subset of these categories
- we also demonstrated the capability of co-training algorithm to improve performance of classifiers