Abusive behavior in social media

Albert-Ludwigs-Universität zu Freiburg

SS 2017

Tarek Saier

Outline

Introduction

- Abusive behavior
- Machine learning
- Natural language processing

Abusive Yahoo! comments

Agressive Twitter accounts

Conclusion

Abusive behavior

Abusive behavior

Abusive behavior

Abusive behavior

Abusive behavior

Abusive behavior

Types

- Frequency (aggressive behavior ↔ bullying)
- Channel (physical, verbal, relational, property)

Offline

- All of the above

Abusive behavior

Abusive behavior

Types

- Frequency (aggressive behavior ↔ bullying)
- Channel (physical, verbal, relational, property)

Offline

- All of the above

Online

- Pysical ✗
- Property ?
- Relational ✓
- Verbal ✓

Abusive behavior

Types

- Frequency (aggressive behavior ↔ bullying)
- Channel (physical, verbal, relational, property)

Offline

- All of the above

Online

- Pysical
- Property
- Relational
- Verbal 🡄

Machine learning

In general

- Unsupervised ↔ supervised
- Classification, regression, clustering, etc.

Our focus

- Supervised classification

Supervised classification

Example

- Classify furniture: chairs ↔ not chairs

Supervised classification

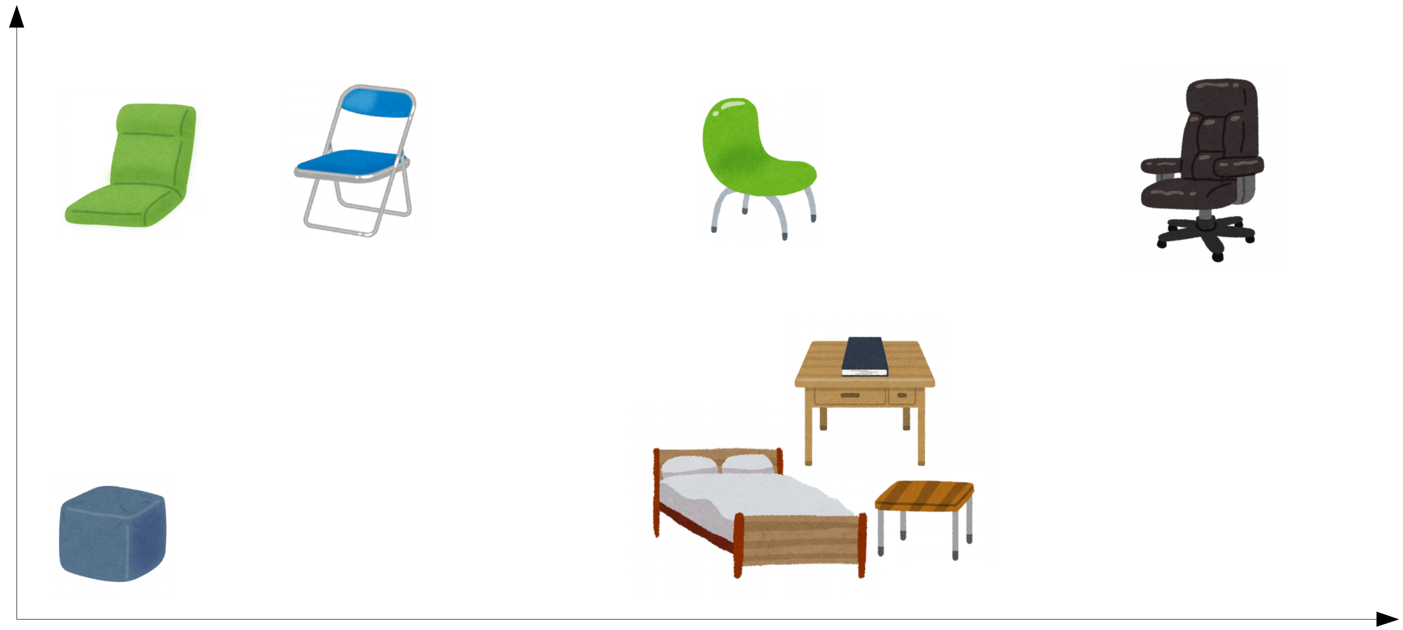
Features

- Number of legs: <int>
- Has backrest: <bool>

Supervised classification

Features

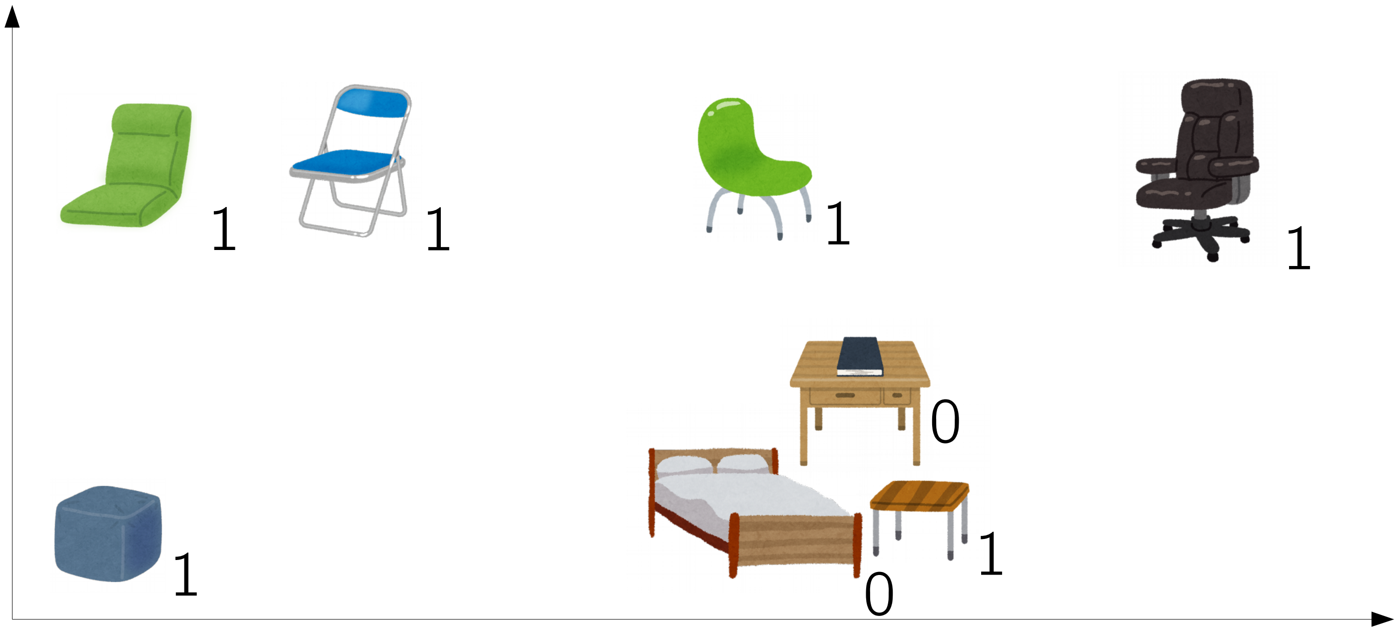
- Number of legs: <int>
- Has backrest: <bool>



Supervised classification

Labels

- Is a char: <bool>



Supervised classification

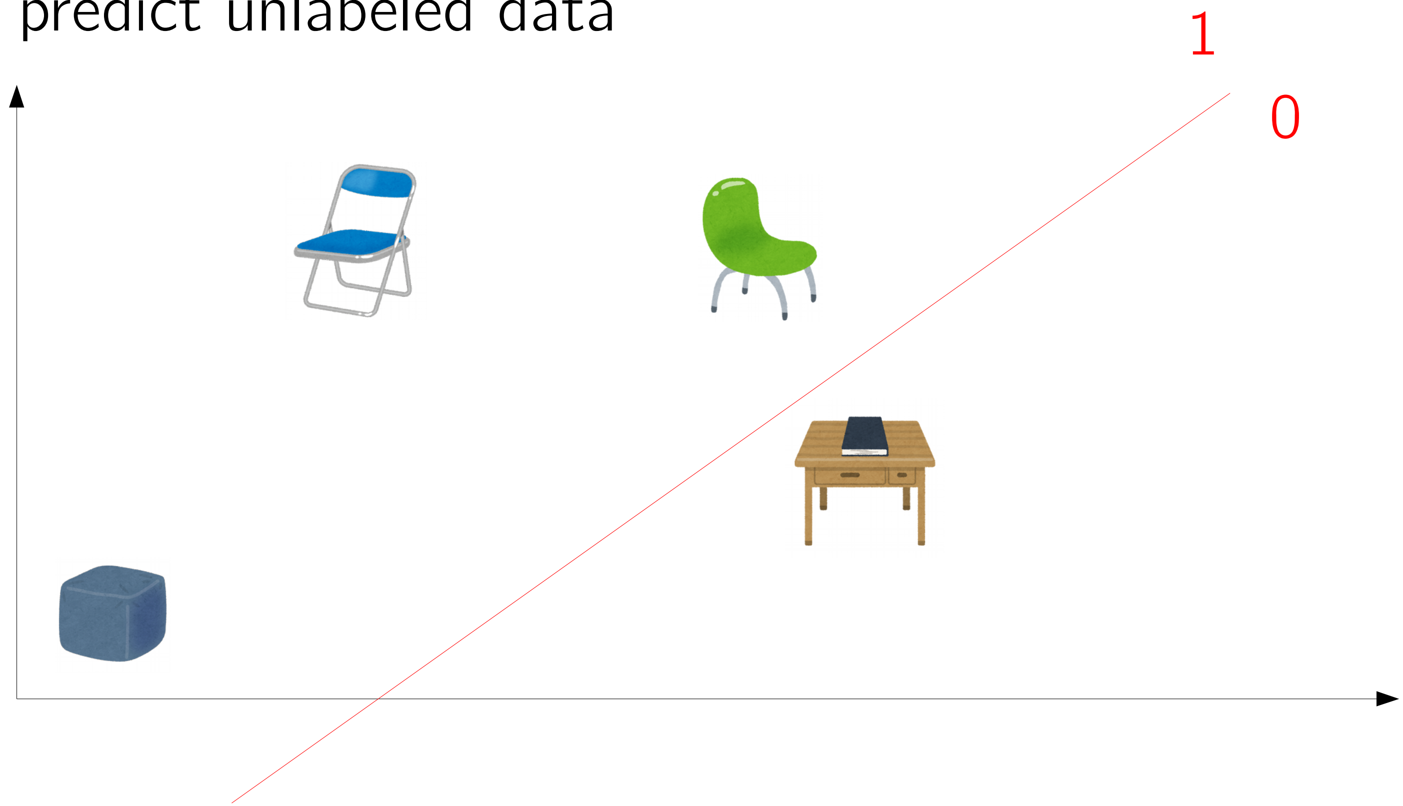
Training / Testing

- often 80/20
- train on labeled data
- predict unlabeled data

Supervised classification

Training / Testing

- often 80/20
- train on labeled data
- predict unlabeled data



Supervised classification

Training / Testing

- often 80/20
- train on labeled data
- predict unlabeled data

Supervised classification

Detecting abusive behavior?

Supervised classification

Detecting abusive behavior?

Social media users

- Activity patterns
- Connections

Textual interaction

- Profanity
- Aggressive language

Natural language processing

Basics

- n-grams
- tf-idf

Advanced

- POS tags
- distributional semantics

Natural language processing

n-grams

- Example: 3-grams $\rightarrow$ Exa, xam, amp, mpl, ple
- Padded: $$P, $Pa, Pad, add, dde, ded, ed$, d$$

tf-idf

- word importance
- $\#\text{term}\big|_{\text{Doc}} \cdot \log \dfrac{\#\text{Doc}}{\#\text{Doc}\big|_{\text{contains term}}}$

Natural language processing

POS tags

- word-category disambiguation
  - noun
  - verb
  - preposition
- ambiguity problem: “The chicken is ready to eat.”

distributional semantics

- ”a word is characterized by the company it keeps”

Natural language processing

Challenges for abusive language detection

- intentional obfuscation
- context dependent acceptability
- context dependent detectability
- ambiguity (sarcasm)
- language evolves

Two concrete approaches

Targets

- Abusive Yahoo! comments
- Aggressive Twitter accounts

Discussion

- Data
- Features
- Experiments

Abusive Yahoo! comments

Detect

- hate speech
- derogatory language
- profanity

In

- Comments
  - Yahoo! News
  - Yahoo! Finance

Abusive Yahoo! comments | Data

Data sets

- primary (2.1 M)
- temporal (1.2 M)
- WWW15 (1 M)
- evaluation (2 K)

Labels

- Yahoo! employees
- Amazon Mechanical Turk*

Abusive Yahoo! comments | Features

Features

- n-grams
- syntactic
- linguistic
- distributional semantics

Abusive Yahoo! comments | Features

Features

- n-grams
- syntactic
- linguistic
- distributional semantics

Syntactic

- inter word dependencies

Linguistic

- average word length
- hate speech words

Abusive Yahoo! comments | Experiments

Primary data set	F-scores	
• all features:	0.80	0.82
• token n-grams:	0.77	0.74
• char. n-grams:	0.73	0.77

Interpretation

- n-grams alone give high quality results

Abusive Yahoo! comments | Experiments

WWW15 dataset

- F-score: 0.78 -

- AUC now: 0.90
- AUC '15: 0.80

Interpretation

- outperforms approach from 2015

Abusive Yahoo! comments | Experiments

Evaluation dataset

- F-score all agree: 0.84
- majority: 0.83
- 2 of 3: 0.43

Interpretation

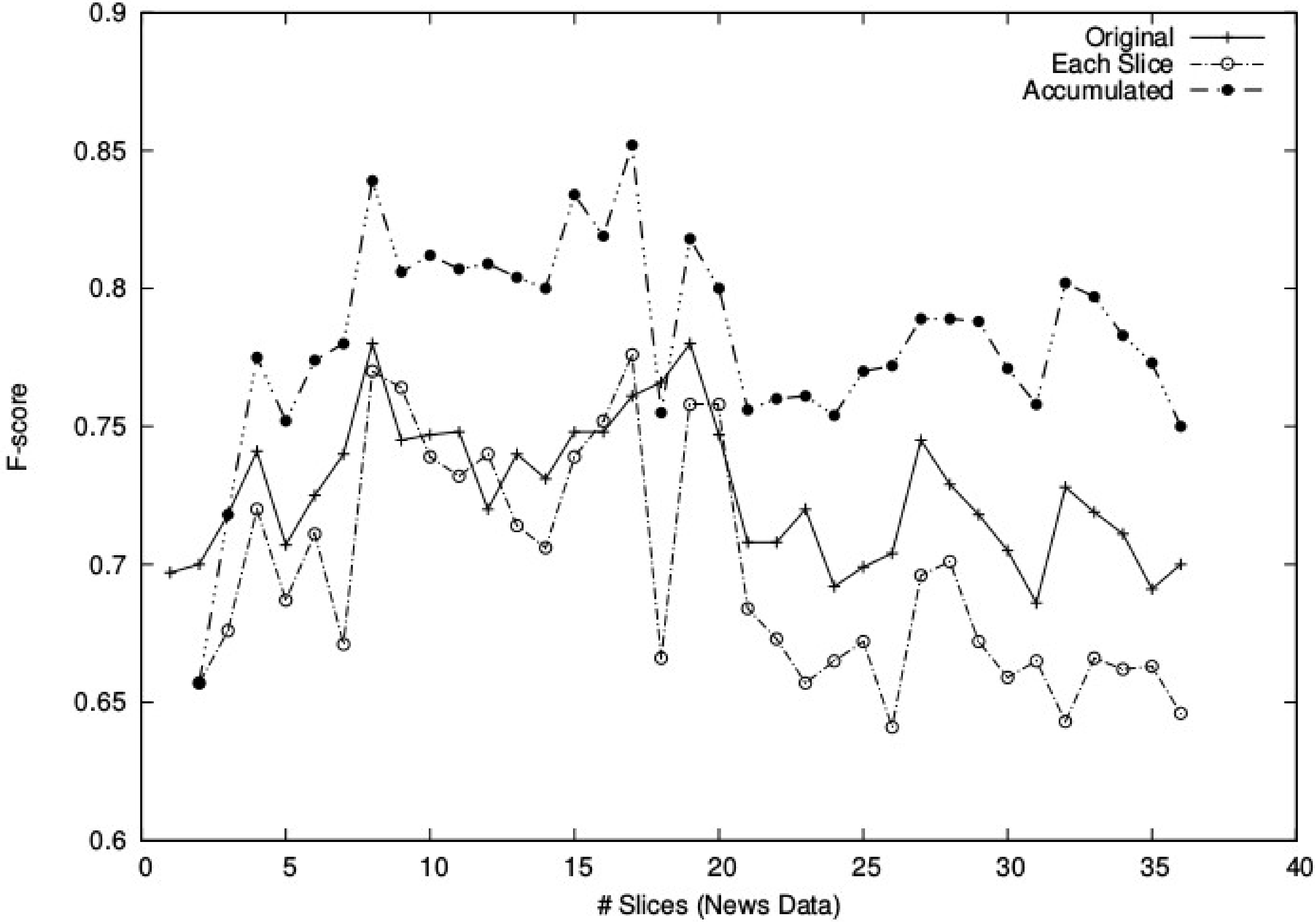
- lanugage is hard to judge

Abusive Yahoo! comments | Experiments

Temporal data set

• Original	train(primary)	predict(t)
• Each slice	train(t)	predict(t+1)
• Accumulated	train(0..t)	preditct(t+1)

Abusive Yahoo! comments | Experiments



Abusive Yahoo! comments | Experiments

Temporal data set

• Original	train(primary)	predict(t)
• Each slice	train(t)	predict(t+1)
• Accumulated	train(0..t)	preditct(t+1)

Interpretation

- recent training data more important than much
- reasonable predictions for small training data

Aggressive Twitter accounts

Detect

- bullying
- aggressive behaviour

In

- Twitter accounts
    - Tweets
    - profile

Aggressive Twitter accounts | Data

Data sets

- baseline (1 M)
- hate related (650 K)
  - #GamerGate

Labels

- CrowdFlower

Aggressive Twitter accounts | Features

User

- account age
- #tweets
- #sessions*

Text

- hate score
- word embeddings

Network

- polularity
- centrality scores

Aggressive Twitter accounts | Experiments

Features

- 12 not useful
  - (session stats, hate score, word embeddings, etc.)

Classification	precision	recall	(in %)
• 4-classes:	72	73	
• 3-classes:	90	92	

Interpretation

- Actual textual content comparably not very useful

Conclusion

Central differences

- abusive content vs. behavior
- target: comments vs. users

Take-home messages

- natural language is hard
- context is valuable

- what about other languages?
- what about non-textual abusive behavior?

Discussion & Questions

"Participation in discussions [...] is also part of the final grade assigned"

(no pressure)