Abusive behavior in social media

Albert-Ludwigs-Universität zu Freiburg

SS 2017

Tarek Saier

Outline

Introduction

- Abusive behavior
- Problem definition
- Challenges
- Solution approaches

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 ⬅

Problem definition

General

- detect (stop) abusive behavior
- here "behavior" = textual communication

Operator's perspective

- available information
    - actor ($\approx$real person/consistent $\leftrightarrow$ anonymous)
- problem modeling
    - abusive user
    - abusive content

Problem definition | Examples

Facebook

- user $\approx$ real person
- friends
- likes
- activity patterns
- ($\longrightarrow$ qualities of user that engages in abusive behavior)

4chan

- consistency throughout one session
- … maybe
- ($\longrightarrow$ abusive content)

Challenges

User

- intentional obfuscation
- sarcasm
  - false positives
  - bullying

Language

- multi sentence
- social context
- language changes

Challenges | Examples

Intentional obfuscation

- $hit
- SHІT
    - cyrillic capital letter byelorussian-ukrainian i (U+0406)
- sh‌it
    - zero width non-joiner (U+200C)

Challenges | Examples

Sarcasm

- false positives:
  "oh of course we should kill them all that would clearly solve all the problems."
- bullying:
  "what a genius idea!"
  "who would've thought of that? amazing!"
  "damn you're smart!"
  "you REALLY know what you're talking about!"
  ...

Challenges | Examples

Multi sentence

- "I tell you the [ethnic group] are the reason. We'd better get rid of them all."

Social context

- utterance X
  - acceptable to group A
  - not acceptable to group B

Language changes

- offensive words loose their impact over time
  $\longrightarrow$ new offensive words

Solution approaches

Word lists

- ✘ blacklists
- ✔ features

n-grams

- example: 3-grams $\longrightarrow$ exa, xam, amp, mpl, ple
- padded: $$p, $pa, pad, add, dde, ded, ed$, d$$
- level: character / token

Solution approaches

tf-idf

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

POS tags

- word-category disambiguation
  - noun
  - verb
  - preposition

$\rightarrow$ similarity

Solution approaches

Domain specific

- URLs
- hashtags
- CAPS
- mentions
- emojis
- etc.

$\rightarrow$ similarity

Solution approaches

Distributional semantics

- word embeddings
  - “the <u>ice</u> cream melted quickly” 0 1 0 0 0
  - “<u>the</u> ice <u>cream melted quickly</u>” 1 0 1 1 1
- word2vec: king − man + woman = queen

Inter word dependencies

- “[ethnic group] are lower class pigs”

$\longrightarrow$ context

Solution approaches

Beyond just text

- user
  - likes
  - activity patterns
  - profile (picture, age, …)
- network
  - centrality
  - polularity
  - reciprocity
- … if possible

Two concrete approaches

Targets

- Abusive Yahoo! comments [1]
- Aggressive Twitter accounts [2]

Discussion

- Data
- Features
- Experiments

[1] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang.
Abusive language detection in online user content.
WWW '16, pages 145–153, 2016.
[2] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali.
Mean birds: Detecting aggression and bullying on twitter.
CoRR, abs/1702.06877, 2017.

Two concrete approaches

Targets

- Abusive Yahoo! comments [1]
    - abusive content
- Aggressive Twitter accounts [2]
    - abusive users

[1] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang.
Abusive language detection in online user content.
WWW '16, pages 145–153, 2016.

[2] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali.
Mean birds: Mean birds: Detecting aggression and bullying on twitter.
CoRR, abs/1702.06877, 2017.

Abusive Yahoo! comments

Detect

- hate speech
- derogatory language
- profanity

In

- Comments
    - Yahoo! News
    - Yahoo! Finance

Abusive Yahoo! comments | Data

Data sets	% abusive	
• primary (2.1 M)	13	
• temporal (1.2 M)	7	
• WWW15 (1 M)	6	
• evaluation (2 K)	50	(tripple labels)

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, POS tags

Linguistic

- average word length, number of CAPS, URLS, tokens, punctuations, 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

Goal and interpretation

- assess feature contribution (only highlights shown above)
- 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

Goal and interpretation

- outperforms approach from 2015 [3]

[3] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati.
Hate speech detection with comment embeddings.
WWW '15, 2015.

Abusive Yahoo! comments | Experiments

Evaluation dataset

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

Goal and interpretation

- test ground truth (calculation) variations
- lanugage is hard to judge

Abusive Yahoo! comments | Experiments

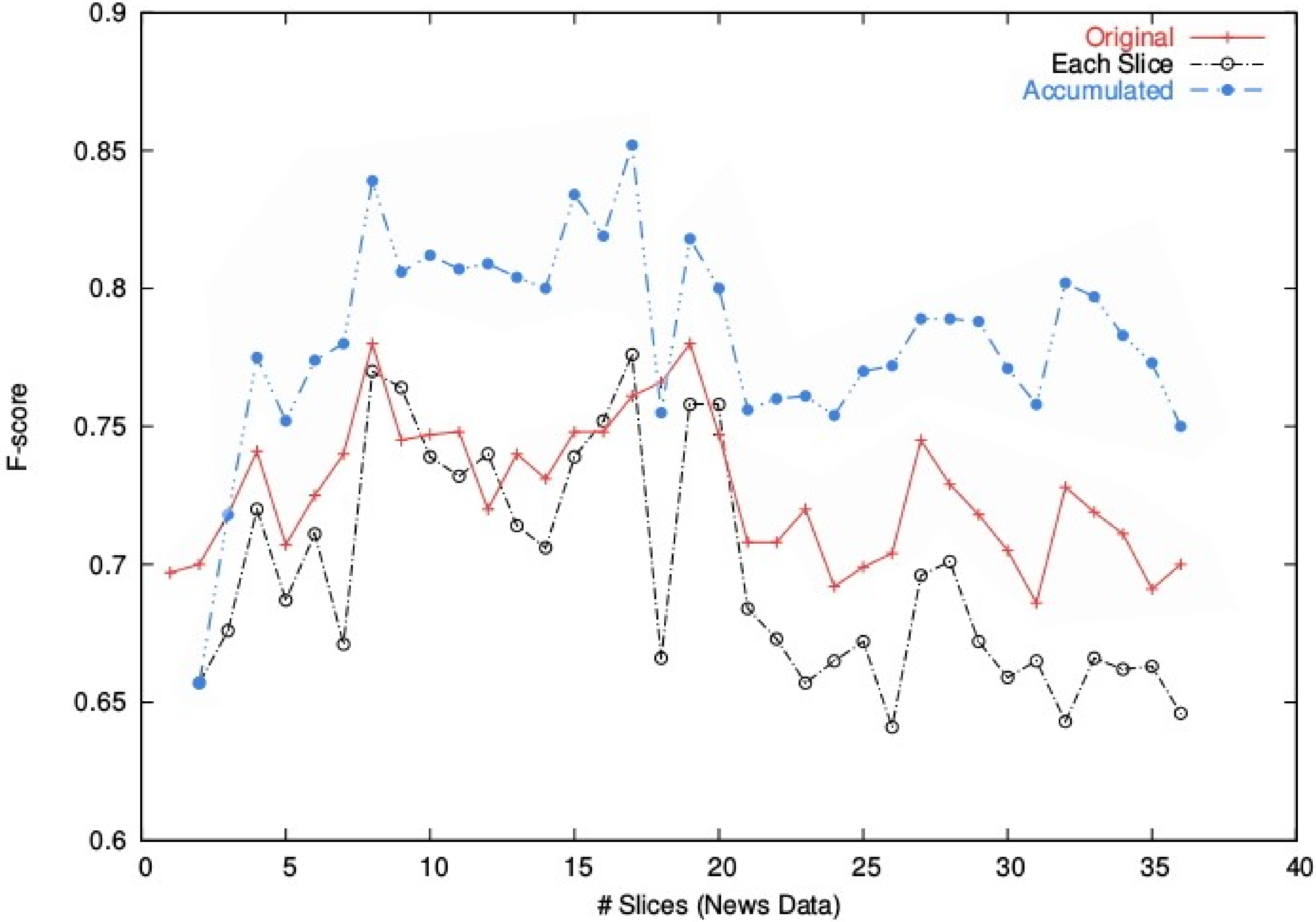
Temporal data set

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

Goal

- assess how much training data is necessary
- assess if updating a model is necessary

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

- Each slice close to Original
  $\rightarrow$ reasonable predictions for small training data
- Accumulated best
  $\rightarrow$ recent training data more important than much

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
- fivefold, majority
- $\rightarrow$ 9,484 annotated
  - 60% normal
  - 32% spam
  - 4.5% bully
  - 3.5% aggressive

Aggressive Twitter accounts | Features

User

- account age, verified,
  interarrival time, num. tweets,
  session statistics

Text

- num. hashtags, emoticons, URLs,
  hate score, word embeddings

Network

- num. friends, followers, polularity (fo/fr),
  reciprocity, centrality scores

Aggressive Twitter accounts | Experiments

Features

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

Classification

- 4-classes: bully, aggressive, normal, spam
- 3-classes: bully, aggressive, normal

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
- approach works well

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)

Sources

- C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang.
  Abusive language detection in online user content.
  WWW '16, pages 145–153, 2016.
- D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, A. Vakali.
  Mean birds: Detecting aggression and bullying on twitter.
  CoRR, abs/1702.06877, 2017.
- N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati.
  Hate speech detection with comment embeddings.
  WWW '15, 2015.
- V. Le Quoc, T. Mikolov
  Distributed Representations of Sentences and Documents.
  CoRR, abs/1405.4053, 2014.