Automatic Classification of Abusive Language and Personal Attacks

in Various Forms of Online Communication

Peter Bourgonje, Julian Moreno-Schneider, Ankit Srivastava, Georg Rehm

peter.bourgonje@dfki.de
DFKI GmbH – Forschungsbereich Sprachtechnologie, Berlin

14. September 2017

http://www.digitale-kuratierung.de

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

WK · POTENZIAL UNTERNEHMEN REGION Die BMBF-Innovationsinitiative Neue Länder

SPONSORED BY THE

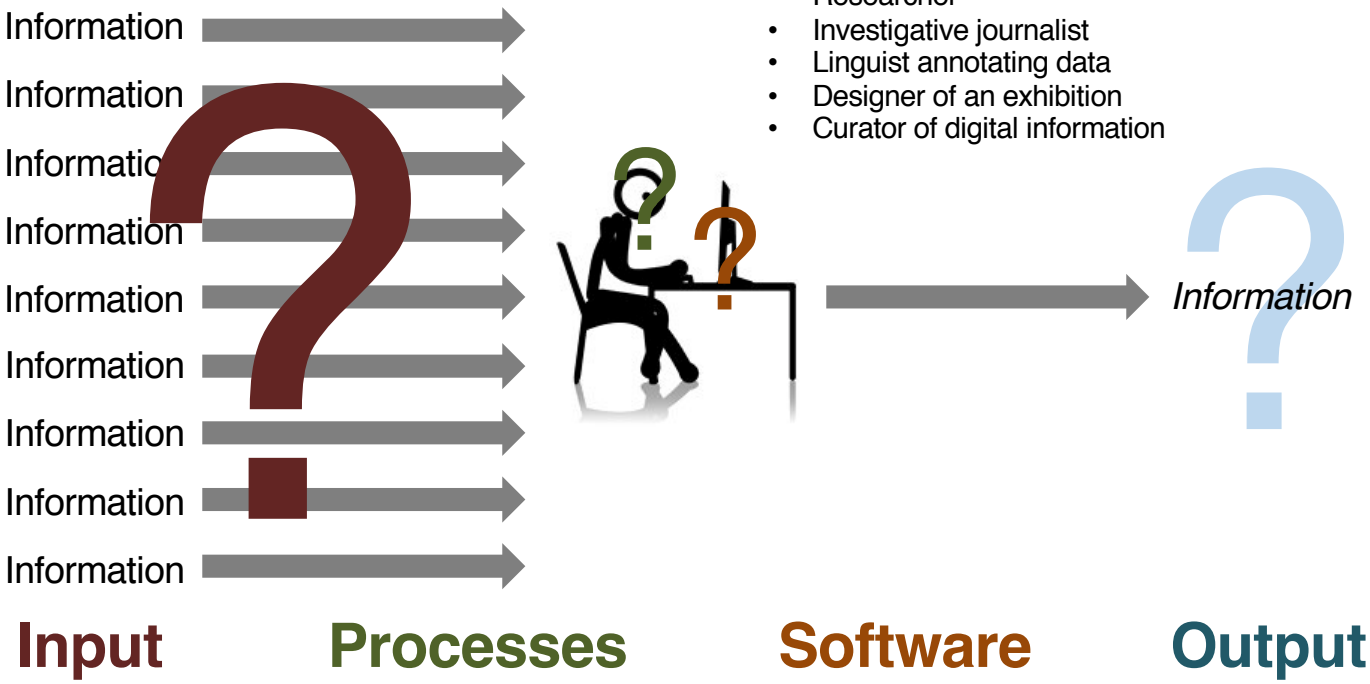
Federal Ministry of Education and Research

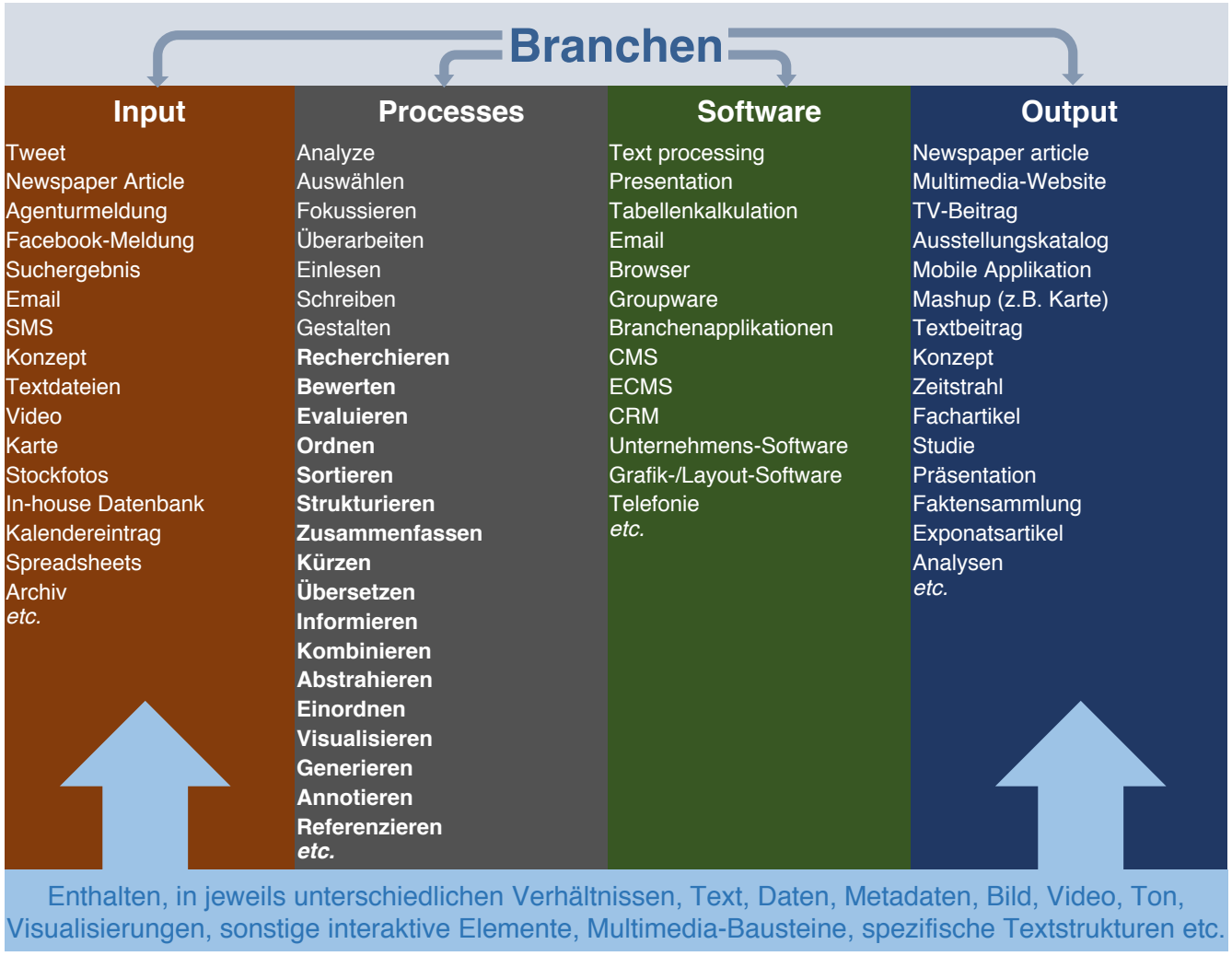
Overview

- Digital Curation Technologies
- Abusive Language & Online Aggression
- Classification Methods and Results
- Conclusions & Future Work

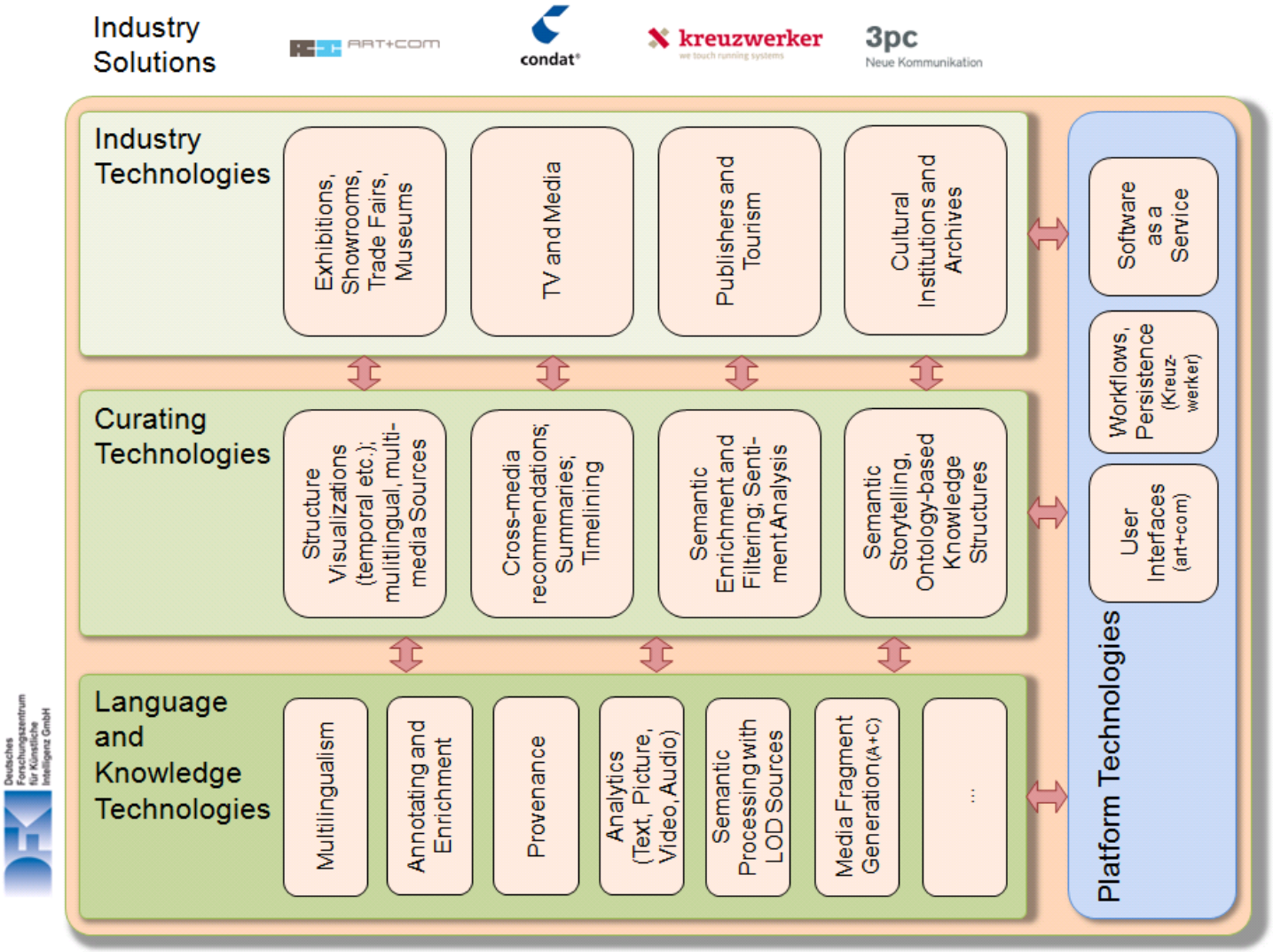
Goals and Use Cases

- Author
- Scholar
- TV editor
- Researcher
- Investigative journalist
- Linguist annotating data
- Designer of an exhibition
- Curator of digital information




Branchen
Input
Tweet
Newspaper Article
Agenturmeldung
Facebook-Meldung
Suchergebnis
Email
SMS
Konzept
Textdateien
Video
Karte
Stockfotos
In-house Datenbank
Kalendereintrag
Spreadsheets
Archiv
etc.
Processes
Analyze
Auswählen
Fokussieren
Überarbeiten
Einlesen
Schreiben
Gestalten
Recherchieren
Bewerten
Evaluieren
Ordnen
Sortieren
Strukturieren
Zusammenfassen
Kürzen
Übersetzen
Informieren
Kombinieren
Abstrahieren
Einordnen
Visualisieren
Generieren
Annotieren
Referenzieren
etc.
Software
Text processing
Presentation
Tabellenkalkulation
Email
Browser
Groupware
Branchenapplikationen
CMS
ECMS
CRM
Unternehmens-Software
Grafik-/Layout-Software
Telefonie
etc.
Output
Newspaper article
Multimedia-Website
TV-Beitrag
Ausstellungskatalog
Mobile Applikation
Mashup (z.B. Karte)
Textbeitrag
Konzept
Zeitstrahl
Fachartikel
Studie
Präsentation
Faktensammlung
Exponatsartikel
Analysen
etc.
Enthalten, in jeweils unterschiedlichen Verhältnissen, Text, Daten, Metadaten, Bild, Video, Ton, Visualisierungen, sonstige interaktive Elemente, Multimedia-Bausteine, spezifische Textstrukturen etc.

Digital Curation Technologies Project



Abusive Language & Online Aggression

- Online media have an unprecedented level of social, political and economic relevance.
- Traditional division between content creator and content consumer is disappearing.
- How to separate high quality content from offensive, hateful, abusive content?

Automatic Classification

- Data sets:
  - English Tweets (Waseem and Hovy, NAACL 2016)
    - 15,979
    - Classes: *sexism, racism, none*
  - German Tweets (Ross et al. NLP4CMC 2016)
    - 469
    - Classes: *hateful* vs. *non-hateful* and on a scale of offensiveness (*1 to 6*)
  - Wikipedia Talk pages (Wulczyn et al. 2016)
    - Subset of 11,304
    - Classes: *attack* vs. *no attack, aggression* vs. *no aggression* and on a scale of aggressiveness *(-3 to 3*)

Examples

- English Tweets:
  - `It's insane they keep bringing people back. When will this show end #MKR` **(none)**
  - `@MKriegbaumJr Yeah! Why do we feed the hungry but not the full? Why do we give shelter to the homeless but not the homed? SO UNEQUAL` **(sexism)**
  - `@FalconEye123456 May Allah bless him with 72 virgin pigs.` **(racism)**
- German Tweets:
  - `Wenn sich der Hass in #Deutschland mal seinen Weg gebahnt hat,wen wird er eigentlich treffen ? #Rapefugees #Politiker #Ossis #Journalisten ?` **(hateful)**
  - `Im #ARD läuft schon wieder #Rapefugees Propaganda.... Zum Kotzen` **(hateful)**
  - `Nordafrikanische Single Männer not Welcome #rapefugees` **(non-hateful)**
- Wikipedia Talk comments:
  - `You stick to your talk page, I'll d mine, right?  20:` **(none)**
  - `== Suck it! == If you can't understand this common American idiom then perhaps you shouldn't be editing Wikipedia.  At any rate, why are you monitoring my talk page, stalker?` **(aggression)**
  - `::::Yes, and Kudpung himself called for an admin's desysop in the section just above this one.  What base hypocrisy.  Perhaps he does not realize his own membership in his "anti-admin brigade", the existence of which he has never provided a shred of evidence for.` **(attack)**

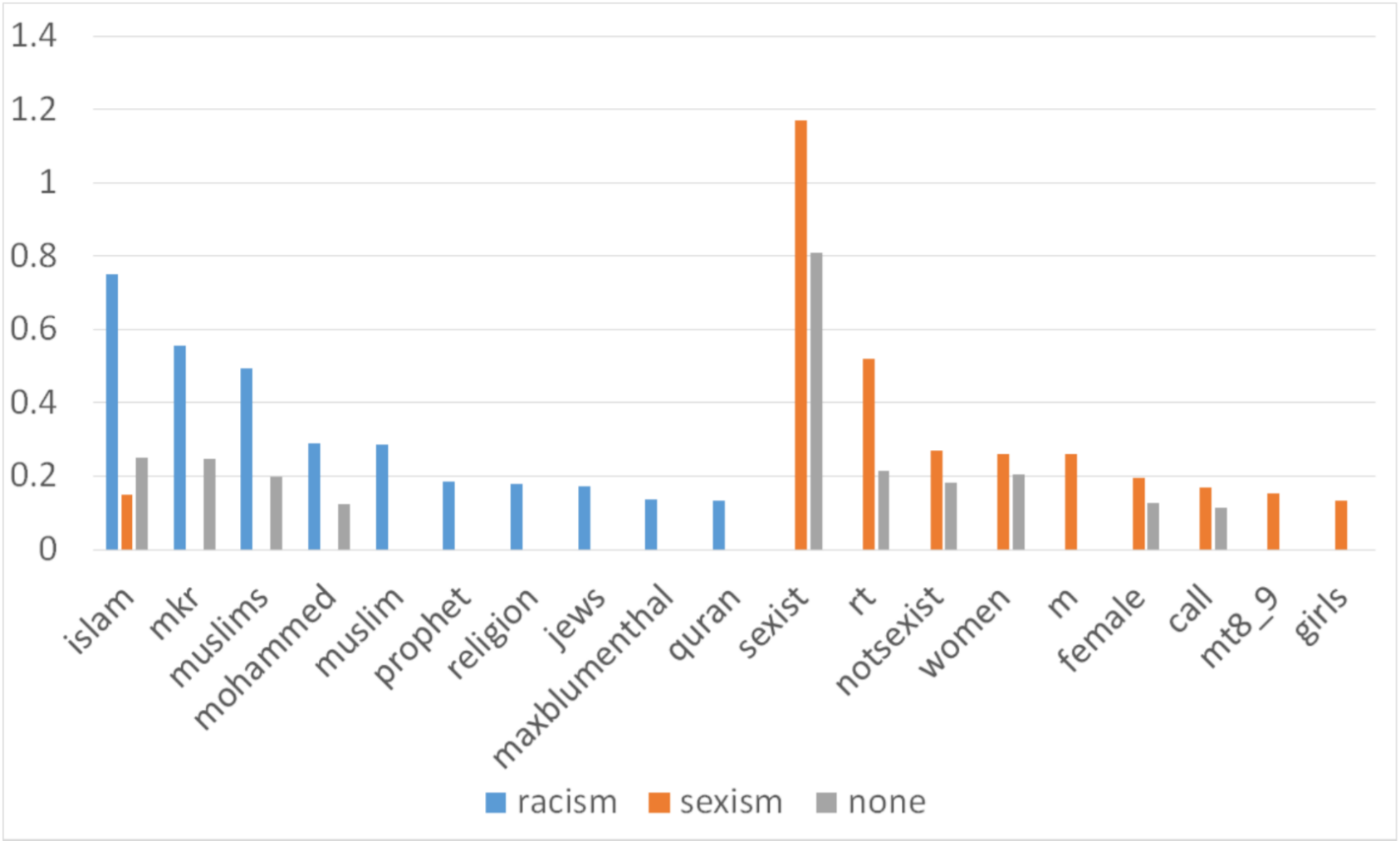
Challenges

- Low inter-annotator agreement values.
- Limited annotated corpora available.
- Accurate detection also dependent on irony and sarcasm, context, author, audience, etc.

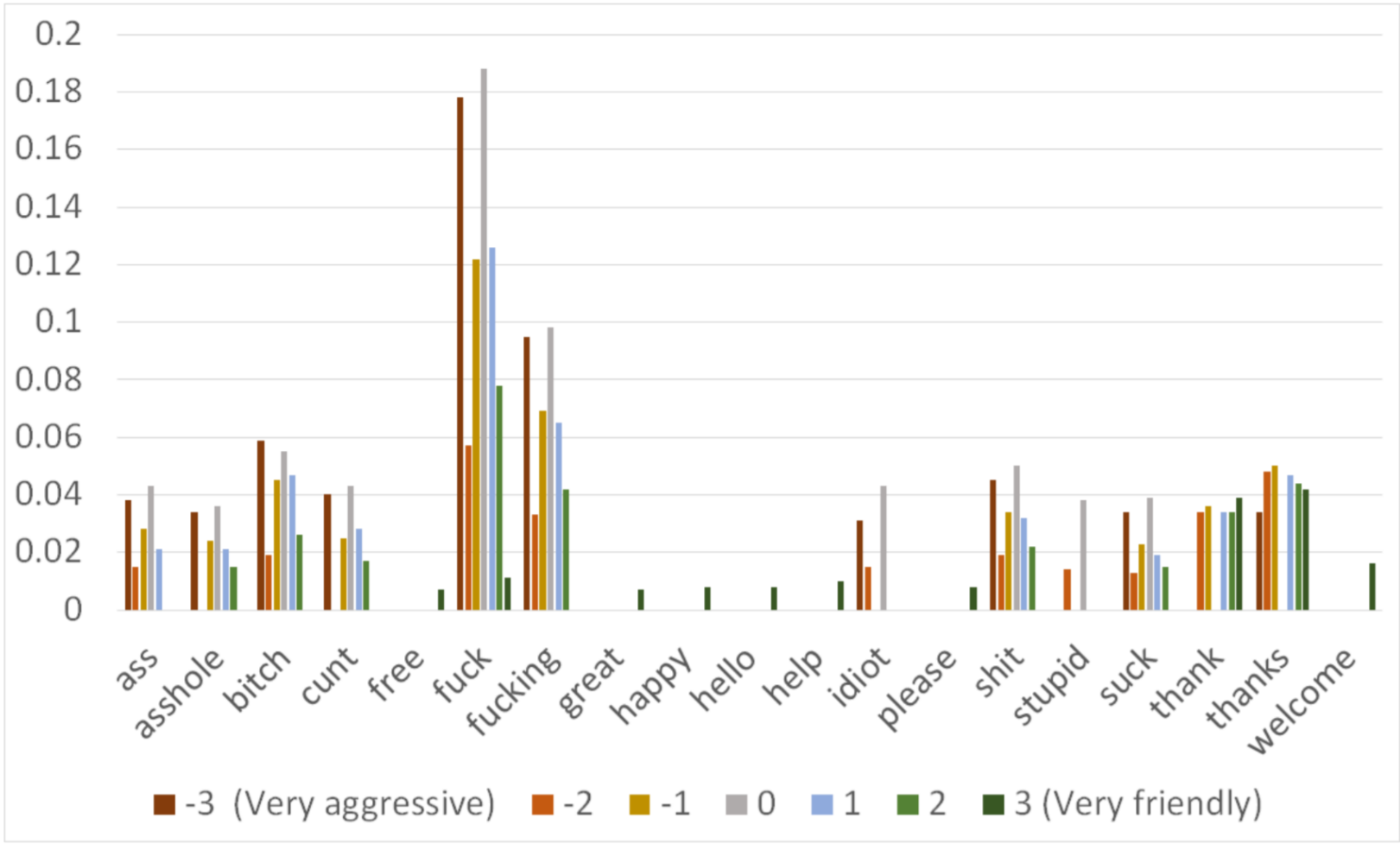
Approach

- Applying a range of classifiers from the Mallet toolkit:
  - Bayes
  - Bayes Expectation Maximization
  - C4.5
  - Logistic Regression
  - Maximum Entropy
  - Winnow2
- All using simple BOW (unigram) features.

	Bayes	Bayes EM	C4.5	Logistic Regression	Maximum Entropy	Winnow2
English Tweets						
Accuracy	84.61	84.01	82.95	**85.67**	83.67	76.66
Precision	80.54	79.57	79.07	**83.57**	81.2	69.85
Recall	**78.63**	77.97	74.37	77.45	74.37	69.62
F1	79.1	78.34	76.17	**80.06**	77.2	69.32
German Tweets (binary, expert 1)						
Accuracy	75.74	**78.93**	74.04	77.23	75.96	71.91
Precision	70.65	**75.07**	69.3	74.8	72.46	72.41
Recall	74.78	76.06	74.98	**76.58**	74.85	72.68
F1	65.84	69.74	70.66	71.98	**73.02**	71.15
German Tweets (binary, expert 2)						
Accuracy	**80.21**	74.26	76.81	79.15	76.38	77.23
Precision	72.76	73.59	72.54	**77.18**	73.62	74.65
Recall	77.57	79.49	77.85	**79.74**	77.31	76.37
F1	70.93	68.97	69.85	**75.41**	74.2	73.05
German Tweets (rating)						
Accuracy	36.6	35.32	**37.87**	33.4	34.89	25.53
Precision	42.51	39.76	**56.22**	31.39	31.9	38.17
Recall	38.53	38.19	**38.76**	36.34	35.71	25.84
F1	27.43	27.03	23.68	30.34	**30.75**	24.06


1.4
1.2
1
0.8
0.6
0.4
0.2
0
islam
mkr
muslims
mohammed
muslim
prophet
religion
jews
maxblumenthal
quran
sexist
rt
notsexist
women
m
female
call
mt8_9
girls
racism
sexism
none

	Bayes	Bayes EM	C4.5	Logistic Regression	Maximum Entropy	Winnow2
Wikipedia Talk (attack binary)						
Accuracy	**83.11**	82.7	81.08	80.9	77.71	77.77
Precision	**81.78**	81.33	79.27	79.36	76.03	77.11
Recall	**83.14**	82.83	81.31	80.97	77.87	77.83
F1	**81.58**	81.36	79.27	79.74	76.65	77.28
Wikipedia Talk (aggression binary)						
Accuracy	**82.19**	82.1	79.58	80.42	77.17	79.08
Precision	**80.68**	80.6	78.13	78.91	75.26	77.25
Recall	**82.01**	81.87	80.18	80.46	77.29	78.57
F1	**80.6**	80.57	78.37	79.23	75.8	77.45
Wikipedia Talk (aggression rating)						
Accuracy	67.13	**67.4**	66.81	65.28	57.77	55.73
Precision	57.21	56.05	54.08	**57.42**	57.21	54.07
Recall	**67.27**	66.94	66.42	65.68	58.18	55.73
F1	59.13	59	58.14	**59.95**	55.26	54.53


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
ass
asshole
bitch
cunt
free
fuck
fucking
great
happy
hello
help
idiot
please
shit
stupid
suck
thank
thanks
welcome
-3 (Very aggressive)
-2
-1
0
1
2
3 (Very friendly)

Conclusions & Future Work

- Reasonable results on these corpora & domains, but:
  - More annotated data needed and from various domains.
  - Inter-annotator agreement remains a serious issue.
  - Inclusion of more sophisticated features from sentiment analysis (scope of abusive words, irony, context).
  - How to model extra-linguistic aspects of abuse and aggression?