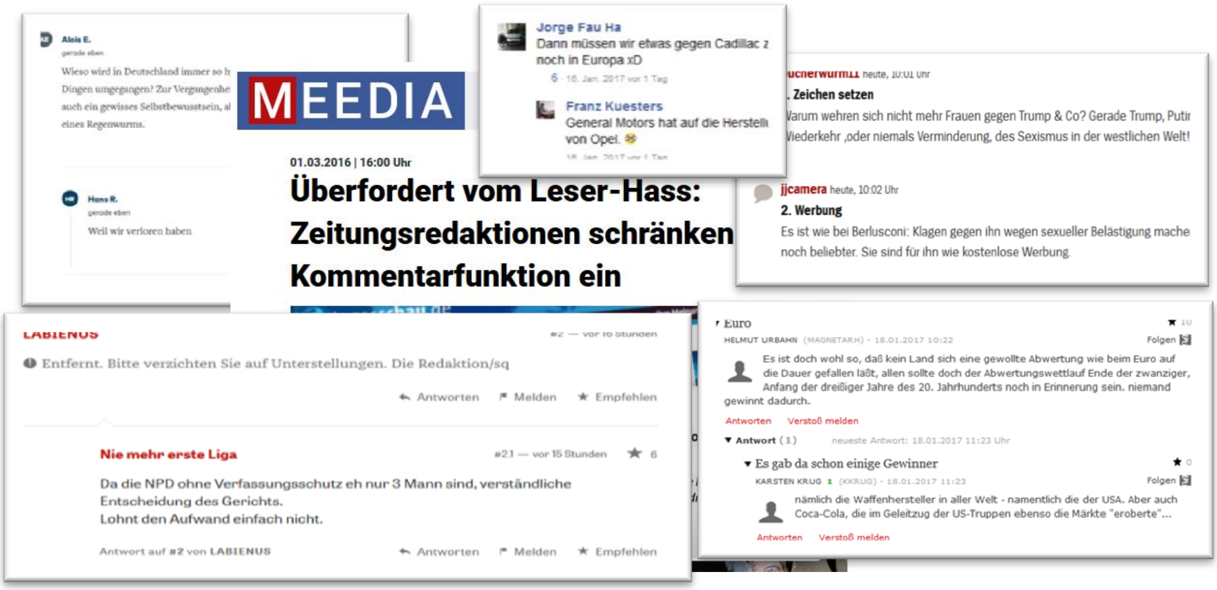
Informationssysteme
Dr. Ralf Krestel
Masterprojekt Sommer 2017

Hate Speech Detection



Aloïs E.
gerade eben
Wieso wird in Deutschland immer so h
Dingen umgegangen? Zur Vergangenhe
auch ein gewisses Selbstbewusstsein, al
eines Regenwurms.

Hans R.
gerade eben
Weil wir verloren haben

MEEDIA
01.03.2016 | 16:00 Uhr
Überfordert vom Leser-Hass: Zeitungsredaktionen schränken Kommentarfunktion ein

Jorge Fau Ha
Dann müssen wir etwas gegen Cadillac z
noch in Europa xD
6 · 16. Jan. 2017 vor 1 Tag

Franz Kuesters
General Motors hat auf die Herstell
von Opel.

ucherwurm11 heute, 10:01 Uhr
. Zeichen setzen
Varum wehren sich nicht mehr Frauen gegen Trump & Co? Gerade Trump, Putir
Viederkehr ,oder niemals Verminderung, des Sexismus in der westlichen Welt!

jjcamera heute, 10:02 Uhr
2. Werbung
Es ist wie bei Berlusconi: Klagen gegen ihn wegen sexueller Belästigung mache
noch beliebter. Sie sind für ihn wie kostenlose Werbung.

LABIENUS #2 — vor 16 Stunden
Entfernt. Bitte verzichten Sie auf Unterstellungen. Die Redaktion/sq
Antworten Melden Empfehlen

Nie mehr erste Liga #2.1 — vor 15 Stunden ★ 6
Da die NPD ohne Verfassungsschutz eh nur 3 Mann sind, verständliche
Entscheidung des Gerichts.
Lohnt den Aufwand einfach nicht.
Antwort auf #2 von LABIENUS Antworten Melden Empfehlen

Euro ★ 10
HELMUT URBAHN (MAGNETARH) - 18.01.2017 10:22 Folgen
Es ist doch wohl so, daß kein Land sich eine gewollte Abwertung wie beim Euro auf die Dauer gefallen läßt, allen sollte doch der Abwertungswettlauf Ende der zwanziger, Anfang der dreißiger Jahre des 20. Jahrhunderts noch in Erinnerung sein. niemand gewinnt dadurch.
Antworten Verstoß melden
▼ Antwort (1) neueste Antwort: 18.01.2017 11:23 Uhr
▼ Es gab da schon einige Gewinner ★ 0
KARSTEN KRUG 1 (KKRUG) - 18.01.2017 11:23 Folgen
nämlich die Waffenhersteller in aller Welt - namentlich die der USA. Aber auch Coca-Cola, die im Geleitzug der US-Truppen ebenso die Märkte "eroberte"...
Antworten Verstoß melden

Comment Analysis

Social media and news provider have to deal with thousands of comments every day. Because many of them violate the terms of service of the platforms, media experts have to manually judge whether a comment is eligible for publication or not. Obviously this is a very sensitive task since no news provider wants to censor valid opinions. Hate speech is one of the content types that need to be filtered out, but there are others as well. From trivial double postings, off-topic comments, to propaganda, and criminally relevant statements, there are all kinds of reasons a comment should not be published.

Filtering Inappropriate Comments

We want to investigate the possibility of helping the experts in finding inappropriate comments before they are published. For this we partner with a large German online news site which provides labeled data of their gathered comments. Data mining and machine learning methods should be used to pre-classify comments. To this end we will train decision trees, naïve Bayes models, support vector machines, and deep learning models[i]. Since a binary decision (good comment – bad comment) is not enough, we also need to provide reasons/explanations for our classification, e.g. "because the comment contains word x and y, it is inappropriate with a confidence of 94%".

Research challenges

- Predict the probability of a given comment to be filtered. An **accuracy** of more than 90% should be achieved
- How does the **context** of a posting influence its classification? Is a comment for one article fine but not for others? What about previous postings and replies to other postings?
- How **time-dependent** is the classification model? Comments that were inappropriate a month ago might be fine today, because of changed political views or events happening, e.g. refugee crisis.

Project Goals

1. Design, Implement, and evaluate a (semi-) automatic system for filtering inappropriate user comments
2. Experiment with multiple machine learning/text mining algorithms
3. Work closely with our project partner
4. Submit a paper describing the task, the system and the developed methods

Contact

Dr. Ralf Krestel
ralf.krestel@hpi.de

[i] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web* (WWW '16). World Wide Web Conf. Steering Committee, 145-153. DOI: https://doi.org/10.1145/2872427.2883062