Hate Speech Detection

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off-topic comments, to propaganda, and criminally relevant statements, there are all kinds of reasons a comment should not be published.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. 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

---