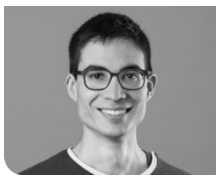
Identifying inappropriate comments in German-language online newspapers

Graduate



Joel Hirzel

Jan Huber

Abinas KUGANATHAN

Objective: On most Swiss news websites, users can comment on articles. Usually, there are commenting guidelines and the comments go through a moderation-process before publication. Inappropriate comments are often referred to as "troll" comments. However, the definition of "online trolling" is ambiguous. Trolling can range from harmless jokes to bullying or state-sponsored propaganda.

To filter inappropriate comments, most newspapers in Switzerland rely heavily on manual moderation. In this thesis, several machine learning models were trained to detect inappropriate comments automatically.

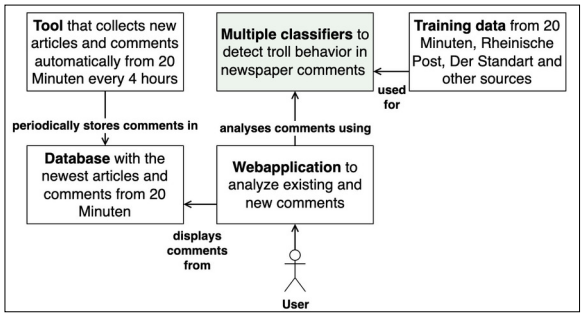
Approach / Technology: Multiple classification algorithms were developed to detect hate speech (84% correct results), off-topic comments (80% correct results) and state-linked propaganda (91% correct results). These classifiers were trained with data from different sources. For comments from the biggest Swiss newspaper, 20 Minuten, the algorithms can correctly predict wether a comment will be accepted or rejected in 70% of the cases.

Result: As a result, the classifiers could be used to support the human moderation team in their work. The performance is not sufficient to fully automate the moderation process – Instead, the algorithms can be used to automatically remove the most extreme comments: By adjusting the threshold, about 20% of the troll-comments can be automatically detected with almost no false positive.

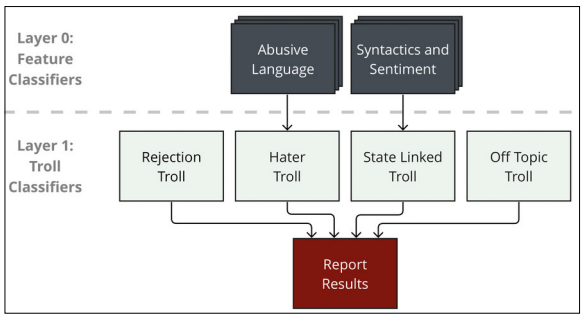
To make the results accessible to lay users, a web application was developed. With the application, users can analyze their own comments with the algorithm or examine existing comments from 20minuten.ch.

In this work, mainly the content of the comments was analyzed. The algorithms are primarily specialized in detecting comments with inappropriate language. For future work, it would be interesting to focus on more subtle manipulation attempts and (paid) political trolls. A challenge in detecting such trolls is the lack of training data. Visualizations and pattern detection could be used to find suspicious patterns. Further research could also focus on the detection of troll accounts instead of individual troll comments.

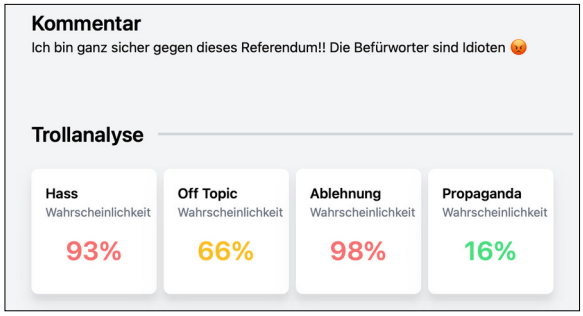
Overview of the project
Own presentment



Overview of the classifiers
Own presentment



Screenshot of the web application showing multiple classifiers
Own presentment



Advisor
Prof. Dr. Daniel Patrick Politze

Co-Examiner
Ramon Schildknecht, SBB AG, Olten 1, SO

Subject Area
Software