Automated detection of inappropriate comments in online discussions

moderate IT

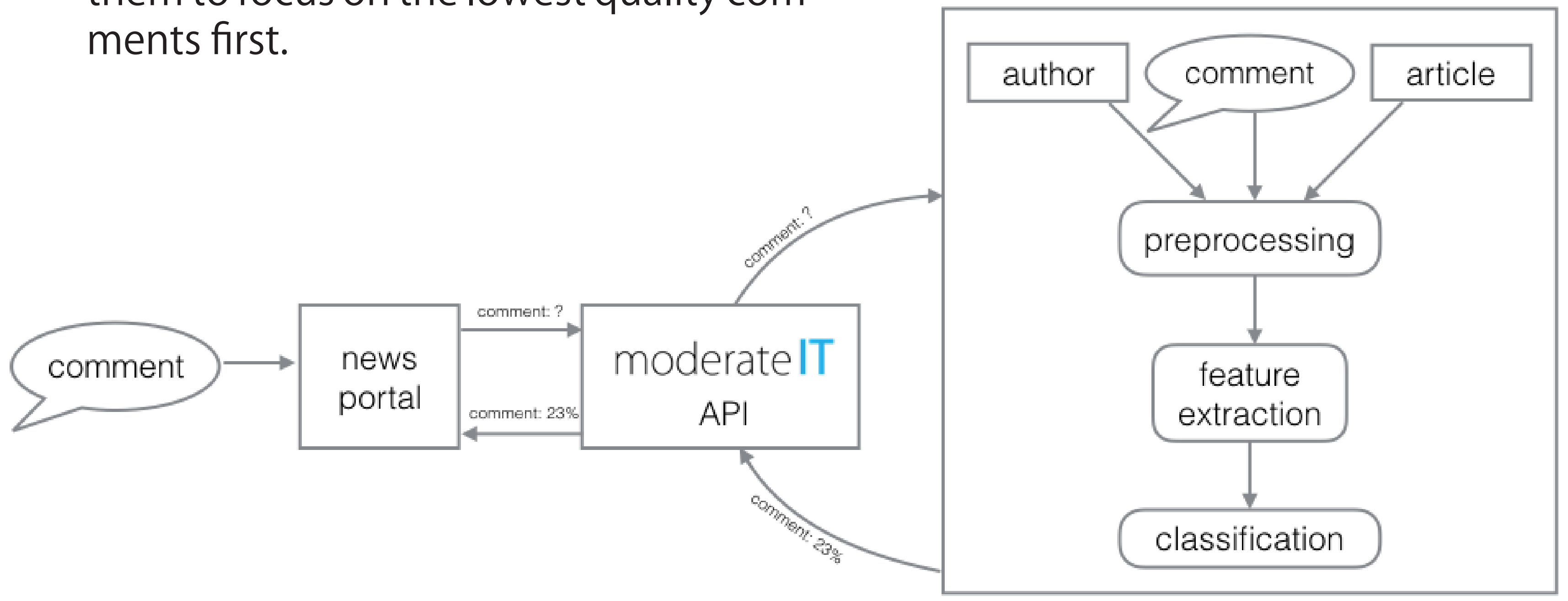
moderate online discussions using IT

Jakub Adam, Monika Filipčiková, Andrej Švec, Filip Vozár
supervisor: Jakub Šimko

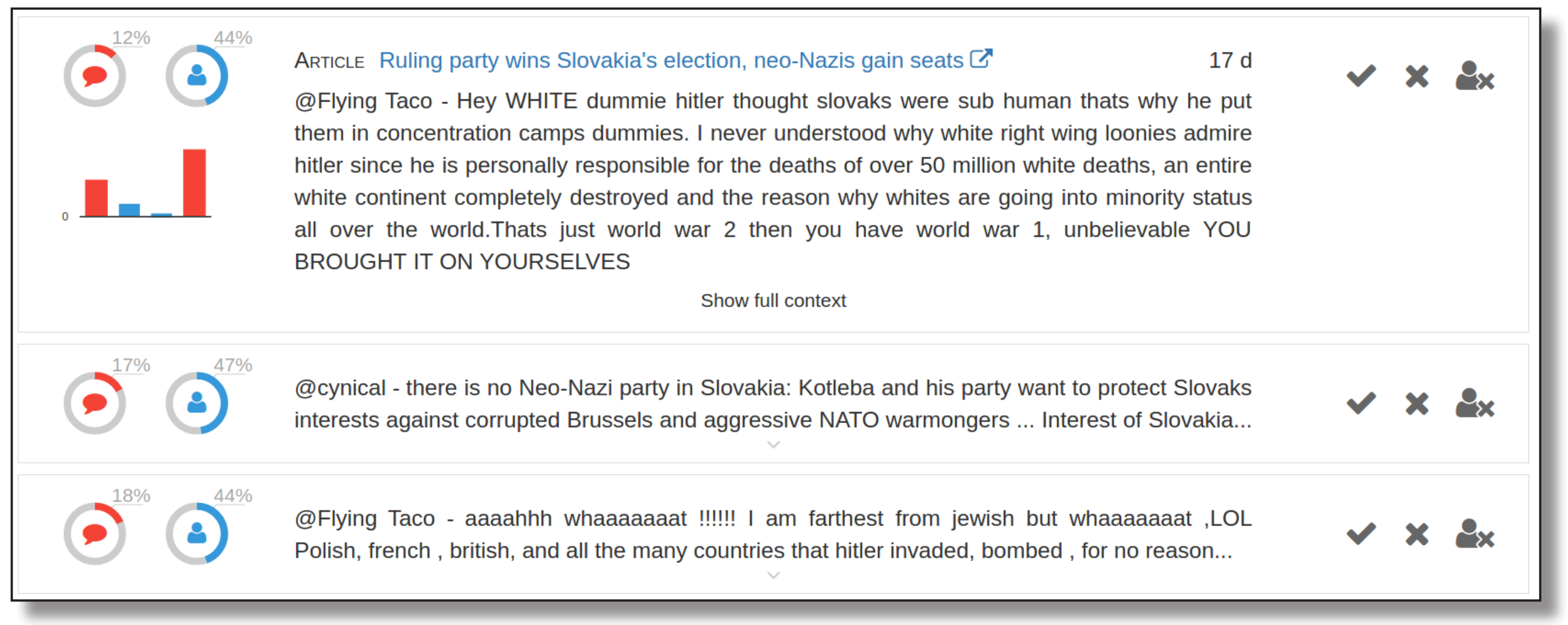
Overview

Our goal is to help online news portals easily manage discussion sections under their articles, which are very often full of hateful and inappropriate comments. Because of the volume of new comments being posted every day, moderators can't check and determine the quality of every comment.

In attempt to solve this situation, we've created moderateIT - a service which can automatically determine comment's quality based on several aspects. Moderators can then sort comments according to our evaluation, allowing them to focus on the lowest quality comments first.



Front-end



Article Ruling party wins Slovakia's election, neo-Nazis gain seats 17 d

@Flying Taco - Hey WHITE dummie hitler thought slovaks were sub human thats why he put them in concentration camps dummies. I never understood why white right wing loonies admire hitler since he is personally responsible for the deaths of over 50 million white deaths, an entire white continent completely destroyed and the reason why whites are going into minority status all over the world.Thats just world war 2 then you have world war 1, unbelievable YOU BROUGHT IT ON YOURSELVES

Show full context

@cynical - there is no Neo-Nazi party in Slovakia: Kotleba and his party want to protect Slovaks interests against corrupted Brussels and aggressive NATO warmongers ... Interest of Slovakia...

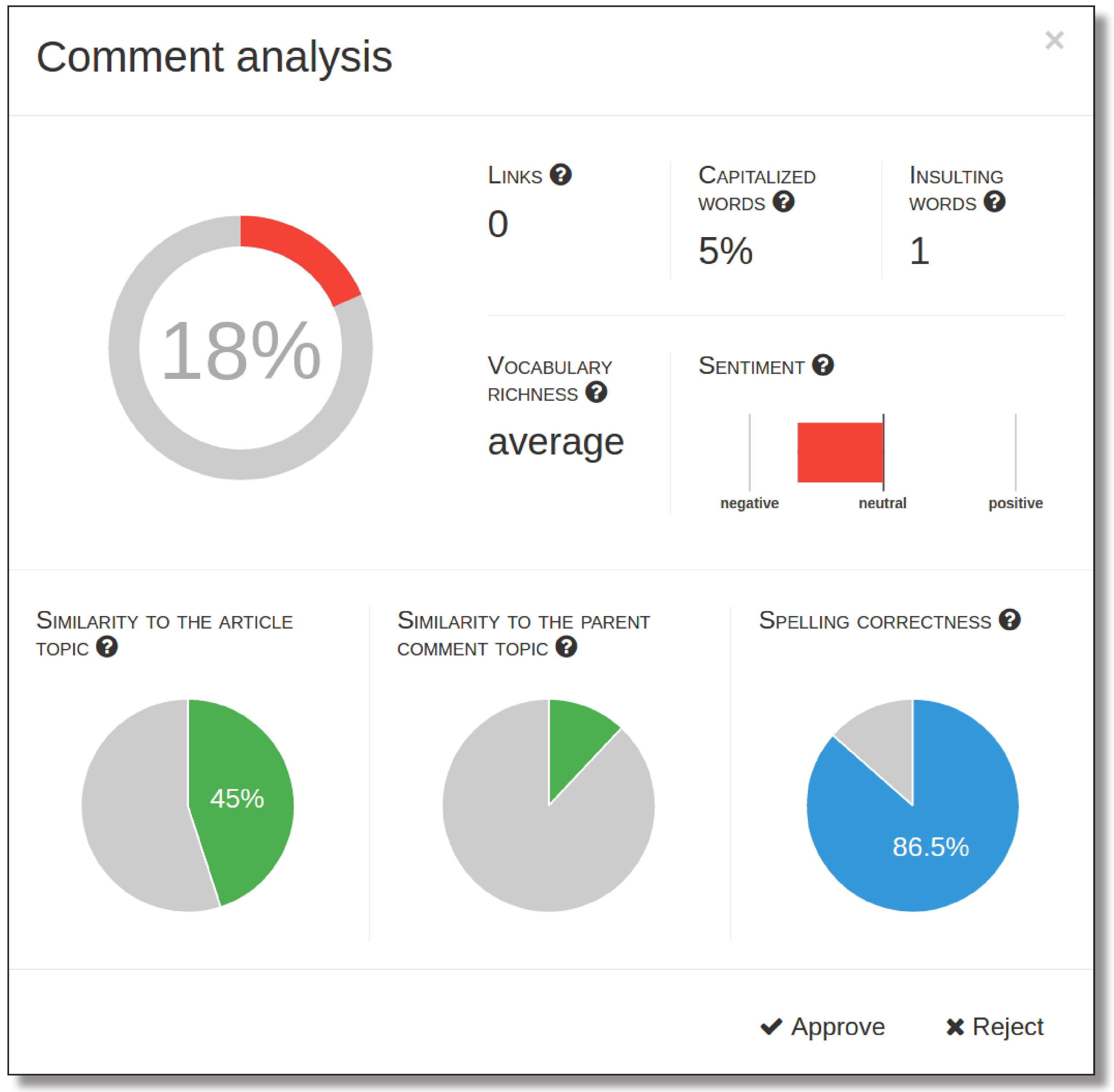
@Flying Taco - aaaahhh whaaaaaaat !!!!!! I am farthest from jewish but whaaaaaaat ,LOL Polish, french , british, and all the many countries that hitler invaded, bombed , for no reason...

Figure 1. Screen with list of comments. They are sorted from worst to best.

We provide a moderation tool for moderators. It is a web front-end built upon our service. This application displays comments on their website along with many features concerning the comment. The moderation task is therefore more efficient. All performed actions are propagated immediately to the news portal. The portal serves also as a presentation tool for Imagine Cup judges or for our future customers.

Comment quality

The analysis process is split into multiple steps. The program runs multiple detectors which identify behavior of each and every comment. Their output then serves as input for classifier and based on these information our program evaluates whether the comment is appropriate or not.



Comment analysis

Links: 0

Capitalized words: 5%

Insulting words: 1

Vocabulary richness: average

Sentiment

Similarity to the article topic

Similarity to the parent comment topic

Spelling correctness

Figure 2. Screen with detailed comment analysis results.

Preprocessing

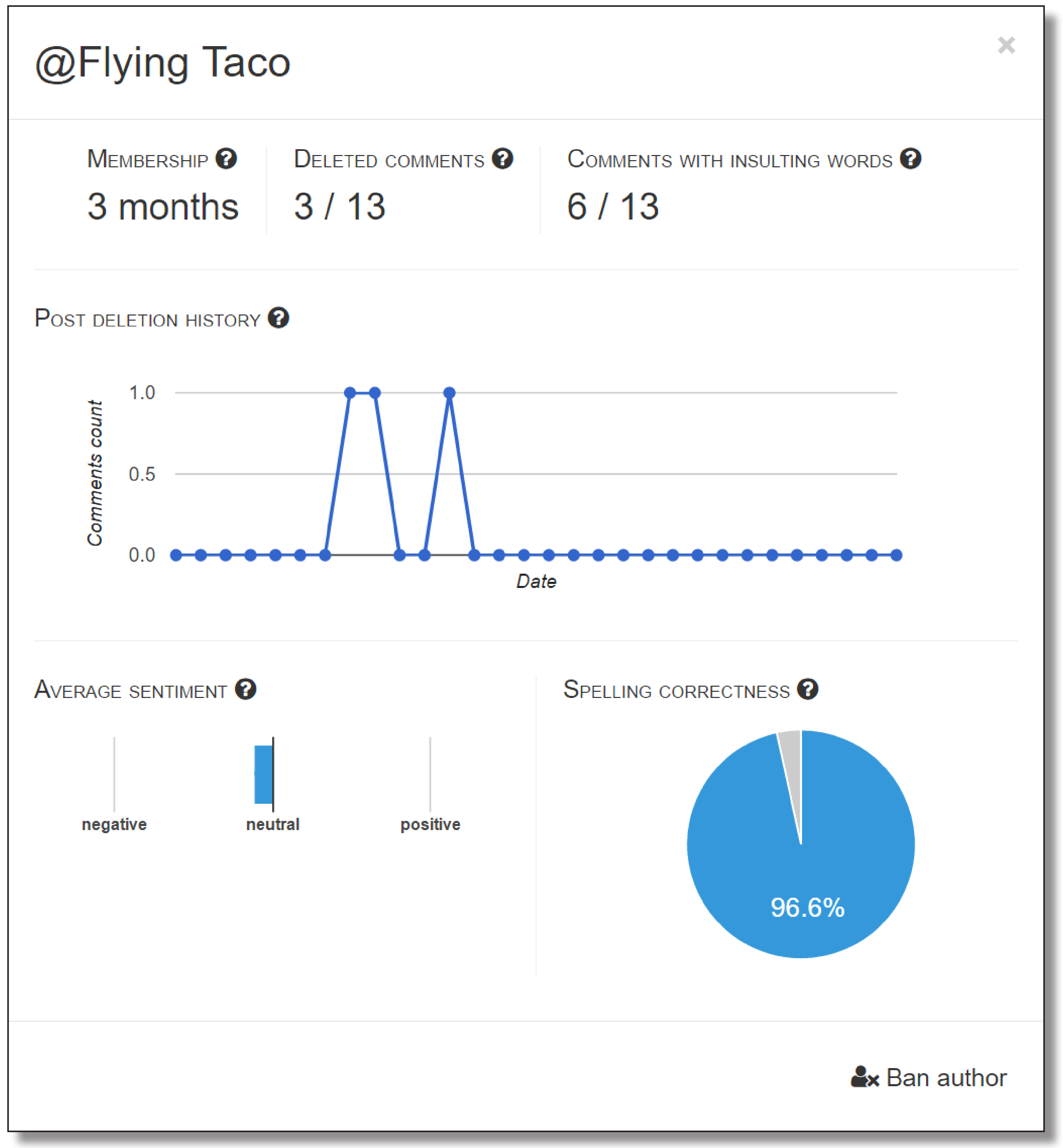
- language detection
- diacritic for slovak text
- lemmatization
- keywords extraction (RAKE - Rapid Automated Keywords Extraction)

Detectors

- insulting words detector
- sentiment detector
- hyperlinks detector
- off-topic detectors
- spellcheck correctness
- vocabulary richness
- emoticons or other non-word characters
- case sensitivity detector
- detector of sequences of the same characters
- formating detector

Author's reputation

User modeling is used for personalisation purpose to improve accuracy of services and products. Most of the discussants have more than one comment in discussions, it is therefore possible to evaluate their history. From long-term point of view trolls write worse comments than usual participants do. We can define values from previous contributions useful for classifier.



@Flying Taco

Membership: 3 months

Deleted comments: 3 / 13

Comments with insulting words: 6 / 13

Post deletion history

Average sentiment

Spelling correctness

Figure 3. Screen with detailed author analysis results. These metrics allow the moderator to see whether author writes insulting comments or whether some of his comments were deleted by a moderator recently, etc.

Features

- duration of user's membership
- changes in user's behaviour
- user's reactiveness to another discussants
- another discussants reactiveness to user
- topic of discussion, where user writes