Delete or not Delete?
Semi-Automatic Comment Moderation for the Newsroom

Julian Risch
Hasso Plattner Institute
University of Potsdam
Prof.-Dr.-Helmert-Str. 2-3
14482 Potsdam, Germany
julian.risch@hpi.de

Ralf Krestel
Hasso Plattner Institute
University of Potsdam
Prof.-Dr.-Helmert-Str. 2-3
14482 Potsdam, Germany
ralf.krestel@hpi.de

Abstract

Comment sections of online news providers have enabled millions to share and discuss their opinions on news topics. Today, moderators ensure respectful and informative discussions by deleting not only insults, defamation, and hate speech, but also unverifiable facts. This process has to be transparent and comprehensive in order to keep the community engaged. Further, news providers have to make sure to not give the impression of censorship or dissemination of fake news. Yet manual moderation is very expensive and becomes more and more unfeasible with the increasing amount of comments. Hence, we propose a semi-automatic, holistic approach, which includes comment features but also their context, such as information about users and articles. For evaluation, we present experiments on a novel corpus of 3 million news comments annotated by a team of professional moderators.

1 Comment Moderation at Online News Providers

Comment sections of online news providers have enabled millions of readers to share and discuss their opinions on news topics publicly. While most people use such platforms to have constructive debates, a tiny minority of individuals or interest groups misuse freedom of speech: by injecting abusive and disruptive comments into online discussions, they spread hate and fear. Furthermore, malicious users misuse the discussion platform to disseminate misinformation with the intent to mislead and provoke readers: fake news. The moderation of online discussions is a huge effort for providers — and the only known way to prevent such attacks, to watch the compliance of users ("netiquette"), and to keep up good discussions.

The definition of inappropriate content is by no means clear and differs between different venues. Many platforms have individual guidelines that users must adhere to and that moderators employ for content assessment. Even with these guidelines there is no precise boundary between appropriate and inappropriate comments. Not only obviously unlawful content, such as ethnic or racial slurs needs to be removed, but the range is wider: personal attacks against other users and the editors, profanity, spam, or off-topic conversations need to be detected and moderated. In addition, legal liability is an issue for news providers and forces them to take action. Given the difficulty of the task, human moderators are needed to manually check and possibly remove comments. Further, it is crucial for moderators to not give the impression of censorship. This is especially the case if opposite positions on emotionally charged issues are involved. Therefore, moderators have to make a difficult choice for each and every comment: *delete or not delete*. With the increasing amount of comments also the costs for moderation increase and manually checking and editing of user-submitted content becomes more and more unfeasible. As a consequence, many large online media sites worldwide were forced to close their discussion areas or downsize them significantly (prominent examples of the last years are Bloomberg, the Internet Movie Database, and the US-American National Public Radio).

Based on a collaboration with a large online news provider, we propose a semi-automatic approach for comment moderation in order to assist human moderators. To this end, it is important to deeply understand the moderation process in the newsroom: moderators take notice of a potentially violating comment

if an attentive user flags the comment as offensive. Besides that, moderators follow discussions in the comment sections preventively and check comments of newly registered users by default. In the event of an inappropriate comment, the moderator can either partly edit or completely delete the comment. Deletion is inevitable if a comment loses its original statement without the violating parts. To ensure a transparent process, moderators leave a message that explains the specific reason for moderation. Users who repeatedly post inappropriate comments can be banned temporarily or permanently. Moderators can close down an article's comment section entirely, if not only a single comment but the majority of comments is inappropriate or the discussion gets off the subject. Although this means can only be the last resort, selected articles are published even without a comment section at all. This shows how severe the problem of moderating inappropriate comments has become for news sites and, unfortunately, how individuals or interest groups successfully attack freedom of speech.

The guidelines of the collaborating news provider list the following attributes for inappropriate comments: (1) insults, discrimination, and defamation, (2) unverifiable suspicions and insinuations that do not rely on plausible arguments or credible sources, (3) advertisements and other commercial content, (4) personally identifiable information of others, (5) copyright reserved texts, (6) quotations without sources, and (7) web links to content that violate these terms of use. All these properties hinder a respectful discussion with other users. On the contrary, a good comment is respectfully worded, argues conclusively, and cannot be misunderstood. It refers to the news article or the subject of discussion.

In order to meet the requirements of the broad definition of inappropriate comments, this paper explores a large set of features. In a holistic approach, we combine information about comments, corresponding articles, and users in a logistic regression model. While we could have used a deep learning approach, such black-box models do not fulfill the requirement of comprehensible classification results. Moderators and readers both need to know the reasons for a classification result. As an advantage of our model, we can give insights into how each feature influences the classification and which features make a comment inappropriate in a specific context.

2 Related Work

The idea to automatically classify insulting messages with a rule-based system goes back to Spertus (1997). A decision tree is trained on a hand-written set of rules with syntactic and semantic text features serving as a basis. Despite the fact that surface-level features, such as bag-of-words, for the most part ignore sentence syntax and even word order, they have strong predictive power (Schmidt and Wiegand, 2017).

Davidson et al. (2017) compared several different classifiers for hate speech detection. Out of logistic regression, naive Bayes, decision trees, random forests, and support vector machines (SVMs), the authors conclude that logistic regression and SVMs perform best. We can confirm this observation with our experiments. In a multi-level approach, Razavi et al. (2010) combine several classifiers with an underlying dictionary of abusive and insulting phrases.

However, small datasets and different labeling schemes currently limit research progress in the field of harassment detection in social media and comment streams (Kennedy et al., 2017). Wulczyn et al. (2017) address this issue by providing machine-labeled data. To this end, a classifier is first trained on a small set of human-labeled comments and afterwards used to generate a larger machine-labeled dataset.

More recently, approaches focus on capturing the semantics of sentences. Methods from word sense disambiguation (Warner and Hirschberg, 2012) and subjectivity detection (Gitari et al., 2015) are explored to detect hate speech. With the rise of distributed dense vector representations, hate speech detection is tackled using deep learning methods. Djuric et al. (2015) propose to learn distributed low-dimensional representations of comments in order to use them as a feature for logistic regression. Nobata et al. (2016) give a comprehensive overview of the various types of abusive language and distinguish three main classes: *profanity*, inappropriate language and swear words; *derogatory speech*, personal attacks; and *hate speech*, directed towards a particular ethnic or religious group. According to the authors, intentional obfuscation, fluency and grammatical correctness of hate speech, make its detection a difficult challenge. The paper explores a variety of features, such as n-grams, linguistic features (length of

comment or average word length), syntactic features (part-of-speech tags), and distributional semantics features. We build on their set of features and extend their approach for the similar, broader task of comment moderation.

Several authors suggest to distinguish separate subclasses of hate speech based on attributes of the attacked groups, such as race, religion or ethnic origin (Warner and Hirschberg, 2012; Gitari et al., 2015). Warner and Hirschberg (2012) observe that hate speech aimed against a specific group often exhibits a high frequency of stereotypical words. Therefore, they create individual language models for each attacked group and determine features, such as word unigrams, separately.

However, word-based models are prone to out-of-vocabulary issues. To this end, Schmidt and Wiegand (2017) propose to generalize exact words to more abstract concepts. They first create clusters of words and then use cluster IDs as features. According to the authors, indicators for appropriate or inappropriate comments are the number of positive verbs, positive adjectives and politeness rules (e.g., "no thanks", "please", "would you"). Polarity classifier for short texts, such as SentiStrength (Thelwall et al., 2010), can be used to count the number of positive, negative, and neutral words in a comment. Our approach makes use of such a polarity classifier and uses sentiment scores as an input feature.

Similar to our approach, Pavlopoulos et al. (2017a) propose a semi-automatic system to assist human moderation teams rather than to replace them. The system tries to classify comments as abusive or non-abusive and if the classification is uncertain a human moderator makes the final decision. Other work of the same authors examines how deep learning with attention layers can be used to moderate user comments and applies a recurrent neural network architecture (Pavlopoulos et al., 2017c; Pavlopoulos et al., 2017b). This approach allows a machine, such as a deep neural network, to delete a comment if its classification is certain enough. In our scenario, a machine that deletes comments automatically but cannot give an comprehensible explanation for its decision is unthinkable.

Park and Fung (2017) compare one step and two step classification. They apply a convolutional neural network as a multi-class classifier in order to detect abusive language. Similar research is conducted by Gambäck and Sikdar (2017), who detect abusive language and also distinguish finer-grained subclasses. Their experiments show a comparable performance for both methods. Finer-grained subclasses can help to give reasons why a comment is classified as abusive.

Napoles et al. (2017b) focus on the complementary task and find "engaging, respectful, and informative conversations". Similarly, Kolhatkar and Taboada (2017) propose to identify constructive comments. Their metric of *constructiveness* (relevant remarks, specific points, and appropriate evidence) is set in contrast to *toxicity* (hate speech, verbal abuse, offensiveness).

Most prior work in the domain of comment classification focuses on English-language datasets. Further, the existing body of work often tackles only a subset of hate speech, such as attacks based on ethnic origin or a specific domain, such as Twitter (Park and Fung, 2017; Badjatiya et al., 2017). In this work, we combine these different approaches and apply them to holistic comment moderation. Our exemplary scenario deals with comment sections at a large German online news provider.

3 Dataset

Our dataset consists of all comments published at a large German online news provider between January 1st, 2016 and March 31st, 2017. In total, there are about 3 million comments by 60k users associated with 26k articles out of which 100k are marked as inappropriate. Each comment is annotated with several moderation flags, which have been manually curated by professional moderators working at the news provider. Based on these flags, we create the binary ground truth for each comment, representing whether the comment is inappropriate or not.

Figure 1 visualizes that the amount of inappropriate comments varies over time, especially due to special or unforeseen high-impact events. The share of inappropriate comments varies between roughly 2% and 10%. We indicate events possibly causing the temporary changes with labels in the figure. For example, terror attacks typically result in emotional and controversial debates, which are prone to include provocative inappropriate remarks. It is worth noting, though, that the general increase in comments following those events might change the way moderators work. Under stress, moderators might choose

Figure 1: The share of inappropriate comments (black) aggregated with a 4-day centered moving average stands out at the date of specific news events. The trend of the total number of comments is shown downscaled for comparison (light gray).

to follow stricter moderation policies. This decision – wittingly or unwittingly – might result in a higher share of flagged comments, even though they do not seem to be objectively worse than similar comments at a different point in time.

Figure 1 also shows that events that gain the most attention are related to social, political or security issues. This circumstance is also mirrored in the news categories with the highest share of inappropriate comments: the categories related to society and politics have a share of 4.1% and 3.4%, respectively, while also containing the majority of all comments (70%). In contrast, the share in all other categories combined is only 2.1%, being lowest in the business category (1.7%). Furthermore, posts are not distributed uniformly among the users: only about 6% of the users posted more than 200 comments, the vast majority of roughly 48% of users have posted only once or twice. As to expect in a real-world dataset that was not meant for academic research purposes, it is far from lab conditions. For example, surprisingly, a single user posted 11,082 comments. 120 of these comments were flagged as inappropriate.

In addition to the platform's guidelines as described in Section 1, we observe that links to foreign language content have been removed by moderators. As not all users can be expected to understand foreign languages, such content hinders them in joining the discussion. Furthermore, moderators remove duplicate comments from the news site, which they do by flagging a duplicate just in the same way as they would do for insults or hate speech. Duplicate detection and hate speech detection are quite different tasks that ask for different approaches. For this reason, we filter exact duplicates in a preprocessing step and resolve data inconsistencies with data cleansing techniques. We aim to detect abusive and disruptive comments that have been moderated because of insults, discrimination, and defamation, but also unverifiable suspicions, which do not rely on plausible arguments or credible sources. These comments hinder a respectful discussion directly. We do not focus on comments flagged because of copyright infringements, web links to inappropriate content, or personally identifiable information.

4 A Holistic, Semi-Automatic Approach to Comment Moderation

We define three different categories of features to classify a given comment:

- Comment features aim to model linguistic, syntactic and semantic properties of the comment's text. Further, comment metadata, such as publication date belong to this feature set.

- User features introduce information about the individuals behind comments and their behavior, in particular the timespan between consecutive posts, previous inappropriate posts, and topics of interest.

- Article features relate to the news article referenced by a comment, for example its category, publication date or article author.

We propose a logistic regression model, implemented with the scikit-learn Python framework[1]. The logistic regression model is trained in a supervised fashion on training data with binary labels. For a given comment with information about the associated user and article, the model predicts a probability of appropriateness. In order to decide about a binary label of appropriateness, we choose a probability threshold that is tailored to achieve a high recall: we want to minimize our false-negative rate.

The set of presumably appropriate comments can be published instantly without any manual contribution. In contrast to that, the set of presumably inappropriate comments is presented to a human moderator for assessment. In practice, a high recall corresponds to the situation where moderators get to see mostly all actually inappropriate comments. The downside of a lower precision is that moderators also need to check a few actually appropriate comments. In a real-world scenario, this trade-off ensures that moderators can be (almost) certain that no inappropriate comment slips through their inspection.

4.1 Comment Features

Comment features are derived from a comment's text or describe a comment's nesting level in the overall thread structure of a discussion.

Linguistic Features Our linguistic features include character-level and word-level features. Neither normalization, such as stemming or lemmatization, nor any other preprocessing is applied. As the most basic feature, we consider a comment's number of characters and number of words. The combination of these two features describes the average word length of a comment. In our dataset, inappropriate comments are on average shorter (48 words) than appropriate comments (61 words). Our other text features focus mostly on extensive use of punctuation or capitalization of whole words. We assume that extensive use of punctuation or capitalization of whole words indicates an aggressive tone.

Comments with web links to external pages frequently violate user guidelines. For example, the linked page's content might contain insults or advertisements. We count occurrences of "http" to capture web links, but do not distinguish between internal and external links. Interestingly, inappropriate comments contain on average less negation words (0.92), such as "not" or "never", than appropriate comments (1.28). In summary, our set of linguistic features for a comment includes: (1) the number of characters, words, and distinct words, (2) the number of question marks, exclamation marks, periods, colons, quotation marks and uses of "http", as well as (3) the ratio of uppercase to lowercase letters.

Syntax Features While the syntax of a comment might not inappropriate itself, it can still serve as an indicator of inappropriateness. For example, the extensive use of personal pronouns might indicate personal attacks against others. Similarly, comments with many adjectives might indicate extensive descriptions or name-calling of other users or of organizations. To capture such behavior, we apply part-of-speech tagging and count the number of adjectives, determiners, personal pronouns, and adverbs in each comment.

Topic Features Off-topic comments that are not related to the article's topic are also considered inappropriate. To measure the topical similarity of an article and a user comment, we apply topic modeling. On a set of roughly 25,000 news articles, we learn a topic model with latent Dirichlet allocation. The topical similarity is then used as a feature in our classifier. Further, we compare the tf-idf vector of the comment with the tf-idf vector of its corresponding article and of inappropriate comments posted at this article. One feature is the cosine similarity of these vectors. Another, similar feature is the Kullback-Leibler divergence of the word frequency distributions of a comment and an article.

Word N-Grams The word cloud in Figure 2b illustrates German unigrams that are most frequently used in inappropriate comments in our dataset. Besides unigrams, we include 2-grams and 3-grams and thereby consider the context in which a word is used. The word cloud shows, for example, that mentioning German chancellor Merkel is already a strong indicator of an inappropriate comment. However, this

[1] http://scikit-learn.org/

indication might change depending on the preceding word: a comment containing "Thank you Merkel" is twice as likely to be inappropriate than a comment containing "Mama Merkel".

Character N-Grams A German-language-specific challenge are arbitrarily long and rare compound nouns. Coining new words extensively increases sparsity of the data: such words typically occur only once in an entire corpus and are called "hapax legomena" by linguists. At test time, this phenomenon leads to frequent out-of-vocabulary problems. Character n-grams do not suffer from such problems, because they are able to capture substrings in compound nouns. Related work shows that character n-grams can be successfully applied to detect abusive language in English-language content (Nobata et al., 2016; Schmidt and Wiegand, 2017). Obfuscated words and unusual spellings typically pose problems for word-based approaches due to a potential of high sparsity, but can be countered with character-based techniques. For this reason, we add character n-grams ranging from length 3 to 5 to our feature set, targeting also the German-specific challenge of compound nouns in particular.

Word2vec To capture the semantic meaning of words, we apply word embeddings. In particular, we use a standard Word2vec approach and model each word as a 100 dimensional vector. The embedding of an article or a user comment is simply the average embedding of all its words.

Structural Features Typical interfaces of discussion sections allow users to post their comments as a reply to another comment. Thereby, a discussion thread can form a tree structure. In our dataset, comments that are direct replies to an article have a higher probability of being inappropriate (4.0%) compared to those that are replies to other comments starting at a depth of 3 (2.5%). Thus, the depth of a comment in the tree, which is the nesting level, is used as a feature.

4.2 User Features

Our dataset also provides information about the user (author) of each posted comment. We define two categories of features based on user information: time-based as well as history-based features. We model four time-based features: the time in seconds since the user last posted an appropriate or an inappropriate comment on the same or any other article. These values are an indicator of heated debates or possible reactions to a previous comment being deleted by the editors, which is frequent in the dataset. For example, if a comment is posted within 10 minutes after an inappropriate comment by the same user on the same article, it has a 19% chance of being inappropriate compared to the global average chance of around 3%.

The history-based features are statistics of a user's comments prior to posting a particular comment. In particular, we count the number of appropriate and inappropriate comments in the same category as the article and globally. Note that since our dataset is limited to the timespan between January 2016 and March 2017, we do not have access to historical information before that time. Therefore, our extracted history-based feature values are only a narrow excerpt of a user's full history. A user who posted only few comments in our dataset might have been much more active in the time before the beginning of our dataset.

4.3 Article Features

Each comment in our dataset is posted in the context of a news article. An article that was just published a few hours ago still has lots of potential for discussion, whereas articles older than a few weeks rarely get new constructive comments. As described in Section 3, the category of the article also influences its probability to receive inappropriate comments. Controversial categories revolving around politics lead to more inappropriate comments than categories about more mundane topics, such as sports. Based on these observations, we define the following features: (1) time since the article's publication, (2) time since the last comment on the article, (3) time since the last inappropriate comment on the article, and (4) category of the article.

Related work has shown that word n-grams are at the top of the best-performing features for hate speech detection (Nobata et al., 2016; Badjatiya et al., 2017; Warner and Hirschberg, 2012; Davidson et al., 2017; Schmidt and Wiegand, 2017). For this reason, we consider word n-grams as a baseline

approach that all other features compete against. Nevertheless, we combine all single features into our holistic approach as a large feature set for the logistic regression. To the best of our knowledge, especially the context of comments, such as information about commenting users and referenced articles, has not been applied so far and extends the state-of-the-art. Our approach could also be used for the sub-task of hate speech detection or related tasks. The broad set of reasons for inappropriateness (as described, for example, in platform guidelines) motivates our large feature set, which might be unnecessary large and might include useless features in other scenarios. Further, other tasks could map probabilities of the linear regression to binary labels differently. If these tasks do not require explanations for automatic decisions, one might refrain from linear models at all in favor of recent deep learning approaches (Napoles et al., 2017a; Badjatiya et al., 2017).

5 Evaluation

We split our dataset time-wise into training and test set in order to train only on past data and evaluate on future data. Stratified sampling is employed to overcome effects of the imbalanced class distribution. The cutoff timestamp is chosen such that 10,000 inappropriate comments remain in the test set. Randomly sampled 10,000 appropriate comments from the remaining comments after the cutoff timestamp are added to the test set. Thereby, a balanced class distribution is obtained in the test set. All comments posted before that timestamp serve as training data. The regression model outputs probabilities of a comment being inappropriate, but not a binary label. In order to be able to compare with our ground truth labels, we map these probabilities to binary labels. To this end, all comments with a predicted probability above a given threshold are marked as inappropriate and all comments with a probability below that threshold are marked as appropriate.

With regard to use cases other than our real-world example, there could be scenarios with two thresholds: one for almost certainly appropriate comments and one for almost certainly inappropriate ones. Only comments between those two thresholds are left for manual assessment. A disadvantage is that the decision to delete a comment could still be made completely automatic if the model is certain enough, which is unthinkable in our scenario.

5.1 Results

Table 2a summarizes the results of our experiments. As can be seen in the table, we choose the threshold in a way that at least 75% of the inappropriate comments are correctly classified, which corresponds to a recall of 75%. The reasoning for this is that it is acceptable to present more comments than necessary to the moderators, but it is important that clearly inappropriate comments do not slip through. On the one hand, if the threshold is set for a higher recall, then the precision decreases until moderators do not profit from machine support but have to check almost all comments. On the other hand, if the threshold is set for a lower recall, more and more actually inappropriate comments are falsely classified as appropriate. As a consequence, inappropriate comments are inadvertently published without intervention.

There are two different error types that can be distinguished. First, our classifier misses to flag an actually inappropriate comment as inappropriate (false negative). Second, our classifier claims to have flagged an inappropriate comment but the ground truth label considers the comment to be appropriate (false positive).

After manual inspection for the first error type it turns out that false negatives are often not clearly inappropriate or further context is needed for the classification. This might be due to the fact that the moderation flags provided in our ground truth dataset have been created by different moderators who interpret the user guidelines differently and therefore make different decisions. Further, different discussions with different preceding comments might ask for more or less intervention by moderators.

As each comment in our dataset has been flagged by at most one moderator, there is no way to evaluate the inter-rater reliability of these flags. Thus, similar comments posted at different times can be flagged differently, which is a tough challenge for classification algorithms. Even a more sophisticated classifier or more features might not help in situations where a team of moderators disagrees at the classification of a particular comment. These hard decisions are another reason why we propose a semi-automatic

Method	Recall	Precision	F_1-Measure
Linguistic	0.773	0.552	0.644
Syntax & Topic	0.757	0.541	0.631
Word N-Grams	0.782	0.559	0.652
Char N-Grams	0.778	**0.649**	0.708
Word2vec	0.756	0.630	0.687
Article	0.753	0.538	0.627
User	0.826	0.516	0.636
Combined	**0.861**	0.615	**0.717**

(a) Overview

(b) Indicative Words for Inappropriate Comments

Figure 2: Classification results and word cloud of indicative words.

approach that integrates machine learning into the manual process of comment moderation.

In the following, we discuss the performance of selected features on our dataset. The analysis of single features is interesting because it helps to understand the result of an otherwise nontransparent automatic classification result. For the support of human moderators, these features can serve as indicators for inappropriate content and can provide explanations for suggested automatic decisions.

5.2 Discussion

Comment features The F1-score of character n-grams is exceeded only by the combined approach feature. Further, it achieves the best overall precision, as listed in Table 2a. One reason for the strong performance on character n-grams might be the complexity of the German language. Word-based models are prone to out-of-vocabulary issues and the German language allows to create very long compound words, which are used infrequently. Character n-grams are more robust than word-level features regarding compound words, neologisms and typos. Some extreme examples for compound nouns in our dataset are "Landesrundfunkanstaltsselbstbedienungsläden" loosely translated as "self-service stores of public service broadcasters" or "Stottertrottelkorbflechterautobahnlastkraftwagenfahrer" loosely translated as "stuttering douchebag who weaves baskets and drives trucks on highways". These nouns are unique in our dataset and also will almost certainly never be used again in any other comment. Even though character n-grams do not capture whole words, they still capture the meaning of a comment well according to our results.

As word embeddings perform comparably well and achieve the second-best F1-score, we further evaluate our approach of calculating a comment embedding by averaging the individual Word2vec embeddings. Therefore, we analyze words that are embedded in the neighborhood of the averaged word vectors. Exemplary results[2] are shown in Table 1. The examples demonstrate that the comment embedding is an adequate way of determining the broader topic of a comment.

Article Features Article features describe the context a posted comment. The classification of a comment as inappropriate only based on the article's category or the time since the article's publication obviously cannot lead to satisfying results. Article features are not a good stand-alone indicator of appropriateness. For example, in order to achieve a recall above 75%, entire categories would need to be considered as inappropriate. Article features seem to be more of a supportive evidence rather than stand-alone features.

User Features An extensive user history is available only for a small subset of all users. A reason for that is the limited time span of the considered dataset. Since the available data covers roughly a year, the observed overall number of comments posted by a user can be only a narrow excerpt of reality. In our dataset, every user initially starts at a comment count of 0, and only later on our model learns

[2] All examples in this paper are translations from German into English.

Table 1: Words in Vicinity of the Comment Embedding

Comment	Undermining democracy: correct, if a court of arbitration is part of a treaty that is an effect of that treaty. Thanks for the tip. DT negotiates, at least according to him, with "America First" in mind. The existing treaties, especially NAFTA, were primarily disadvantageous to Joe Sixpack AND Juan Prez in both Mexico AND the USA.. If Mexico does not like DT's suggestions they can mutually cancel the treaty with the USA and both profit. To rephrase it, it can only get better for the lower middle class and below.
Nearby Words	free trade agreement, trade agreement, trade treaty, treaty, free trade contract
Comment	The Kurdish population is being blackmailed by the PKK and hauled off into the mountains. If the Kurds in the FRG are for the PKK, then they should live there.
Nearby Words	Kurd, PKK, Kurdish area, Turkey, terrorist

to distinguish different users, for example based on the frequency of their comments We expect the features to perform better with a complete user history. Nevertheless, several features show promising results. The number of inappropriate comments by the user in the same article category and in total in the past strongly correlate with a new comment by the user being inappropriate. The time since the last inappropriate comment by the user also weakly correlates with a new comment by the user being inappropriate.

6 Conclusions and Future Work

In this paper, we studied the task of comment moderation. In contrast to the sub-task of hate speech detection, moderation needs to consider several other types of inappropriate comments, such as insults and defamation, but also unverifiable suspicions, which do not rely on plausible arguments or credible sources. Moderators ensure respectful, engaging, and informative discussions by editing or deleting inappropriate comments that do not facilitate constructive debates or even contain unlawful statements. This costly manual task puts a strain on the moderators, forcing more and more news sites to close down their comment sections. In order to meet this challenge, we propose a semi-automatic approach for assisting humans at comment moderation. Our holistic approach combines information about comments, users, and articles in a logistic regression model.

Despite the recent success of deep learning for natural language processing, the lack of comprehensibility of neural nets motivates the application of other models. For example, decision trees and regression models can give reasons for their final prediction results. A particular feature with a particularly large value or a combination of several features might be such a reason. In practice, moderators and users need to understand the automatic decision process underlying a trained machine learning model and in particular *why* a comment is classified inappropriate. For the application in a newsroom and the integration into the moderation process, it is important that classification results are comprehensible.

In future work, we plan to extend the explanatory aspect alongside the classification of comments. Explanations could, for example, include details about the user's comment history or highlight relevant words in the posted text. Such explanations would help moderators to understand the reasons for an automatic decision and explanations make the system and its decisions also more trustworthy. However, provided these explanations, malicious users will certainly use their criminal energy trying to fool automatic moderation approaches.

Acknowledgments

We thank our students Carl Ambroselli, Yannick Bäumer, Andreas Burmeister, Jan Ladleif, and Tim Naumann for their help with this project. We also thank Andreas Loos for his valuable feedback.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the International Conference on World Wide Web (WWW Companion)*, pages 759–760. International World Wide Web Conferences Steering Committee.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 512–515.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 29–30. ACM.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the Workshop on Abusive Language Online*, pages 85–90. ACL.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the Workshop on Abusive Language Online*, pages 73–77. ACL.

Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the Workshop on Abusive Language Online*, pages 11–17. ACL.

Courtney Napoles, Aasish Pappu, and Joel R Tetreault. 2017a. Automatically identifying good conversations online (yes, they do exist!). In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 628–631.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017b. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the Linguistic Annotation Workshop*, pages 13–23.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 145–153.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the Workshop on Abusive Language Online*, pages 41–45. ACL.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017a. Deep learning for user comment moderation. In *Proceedings of the Workshop on Abusive Language Online*, pages 25–35. ACL.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017b. Deeper attention to abusive user content moderation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1136–1146.

John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. 2017c. Improved abusive comment moderation with user embeddings. In *Proceedings of the EMNLP Workshop: Natural Language Processing meets Journalism*, pages 51–55. Association for Computational Linguistics.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Proceedings of the Canadian Conference on Advances in Artificial Intelligence (AI)*, pages 16–27. Springer-Verlag.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of the National Conference on Artificial Intelligence and Conference on Innovative Applications of Artificial Intelligence (AAAI /IAAI)*, pages 1058–1065. AAAI.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology*, 61(12):2544–2558.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Workshop on Language in Social Media (LSM)*, pages 19–26. ACL.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 1391–1399. International World Wide Web Conferences Steering Committee.