Toxic Comment Detection in Online Discussions

Last edited by **Hamed Babaei** in 42 seconds

Toxic Comment Detection in Online Discussions

> Julian Risch and Ralf Krestel, Postdam, Germany, 2019

The misuse by spammers, haters, and trolls makes costly content moderation necessary. Sentiment analysis can not only support moderation but also help to understand the dynamics of online discussions. A subtask of content moderation is the `identification of toxic comments`.

- **A `toxic` comment is defined as a rude, disrespectful, or unreasonable comment that is likely to make other users leave a discussion.**

- in this paper, a fine-grained classification scheme introduced for toxic comments and motivates the task of detecting toxic comments in online discussions.

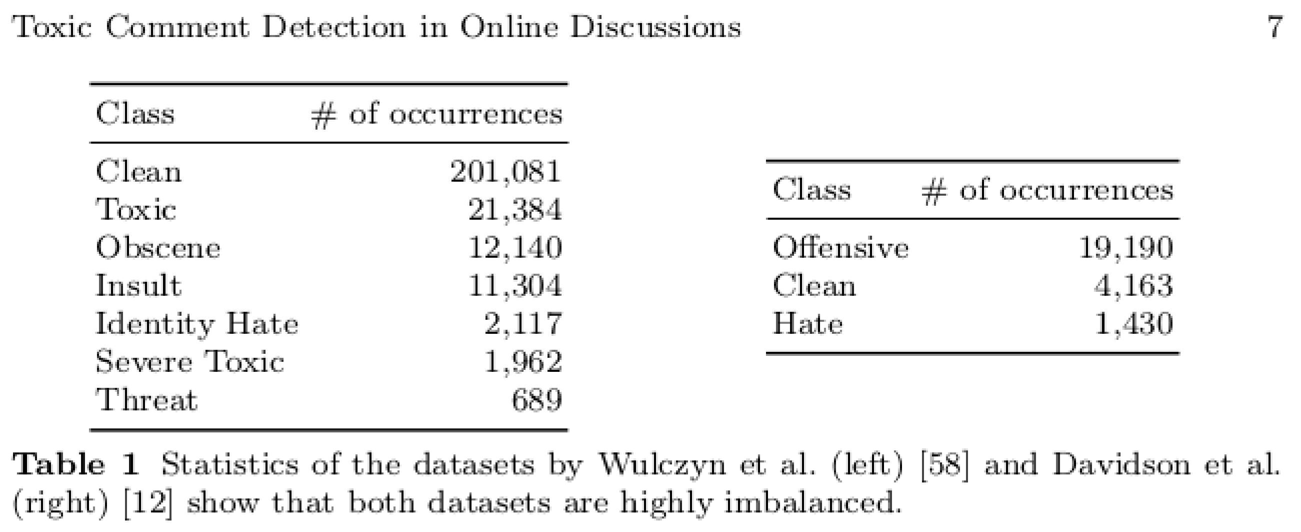
Classes of Toxicity:

generalized/directed and explicit/implicit: `Directed` means a comment addresses an individual, while `generalized` means it addresses a group. `Explicit` means, for example, outspoken name-calling, while `implicit` means, for example, sarcasm or other ways of obfuscation. Other terms for this dimension are overtly and covertly `abusive`. Here are examples of different classes.

- **Obscene Language/Profanity**: This class considers swear or curse words. example: `That guideline is bullshit and should be ignored`
- **Insults**: This class includes statements about individuals or groups. example: `Do you know you come across as a giant prick?`
- **Threats**: Severely toxic comments. Statements that announce or advocate for inflicting punishment, pain, injury, or damage on oneself or others fall into this class. example: `I will arrange to have your life terminated.`
- **Hate Speech/Identity Hate**: In contrast to insults, identity hate aims exclusively at groups defined by religion, sexual orientation, ethnicity, gender, or other social identifiers. example: `Mate, sound like you are Jewish. Gayness is in the air`
- **Otherwise Toxic**: Comments that do not fall into one of the previous four classes will fall into this class. example: `Bye! Don’t look, come or think of coming back!`

Comment Datasets for Supervised Learning

- Yahoo News Annotated Comments Corpus: (522k unlabeled and 10k labeled comments)
- One Million Posts Corpus (1M unlabeled and 12k labeled comments)
- A collection of Wikipedia discussion pages (100k human-labeled and 63M machine-labeled comments)



Toxic Comment Detection in Online Discussions 7

Class	# of occurrences
Clean	201,081
Toxic	21,384
Obscene	12,140
Insult	11,304
Identity Hate	2,117
Severe Toxic	1,962
Threat	689

Class	# of occurrences
Offensive	19,190
Clean	4,163
Hate	1,430

Table 1 Statistics of the datasets by Wulczyn et al. (left) [58] and Davidson et al. (right) [12] show that both datasets are highly imbalanced.

- Davidson et al.[12]: dataset comprises 25k labeled tweets that have been collected by searching the Twitter API for tweets that contain words and phrases from a **`hate speech lexicon`**.

Neural Network Architectures

- **FastText** provide alternative ways to calculate word embeddings. FastText is particularly suited for toxic comments because it uses subword embeddings. The advantage of subword embeddings is that they overcome the OOV problem.

- **Word2Vec** and **GloVe** fail to find a good representation of these words at test time because these words never occurred at training time.
- The ability to cope with unknown words is the reason why previous findings on the inferiority of word embeddings in comparison to word `n-grams` have become outdated.
- RNN layers, such as long short-term memory (LSTM) or gated recurrent unit (GRU) mostly used for toxic comment classifications.
- An extension to standard LSTM and GRU layers are bi-directional LSTM or GRU layers.
- All recurrent layers can either return the last output in the output sequence or the full sequence. it can return the last output in the sequence (a representation of the full input comment) or return outputs of each step in the sequence as an alternative. It is So-called pooling layers that can combine this sequence of outputs. It used to reduce input with many values to an output of fewer values.
  - For toxic comment classification, both average-pooling and max-pooling are common with a focus on the latter.
  - If a small part of a comment is toxic, max-pooling will focus on the most toxic part and finally result in classifying the comment as toxic.
  - In contrast, with average-pooling, the larger non-toxic part overrules the small toxic part of the comment and thus the comment is finally classified as non-toxic Therefore, max-pooling is more suited than average-pooling for toxic comment classification because according to the definition of toxicity classes typically assume that there is no way to make up a toxic part of a comment by appeasing with other statements.
- An alternative to pooling after the recurrent layer is an attention layer that successfully applied to toxic comment classification. [13]
- To prevent overfitting due to the small number of training data dropouts used:
  - Standard Dropout
  - Spatial Dropout
  - Recurrent Dropout

Overview of Approaches and Datasets and Architectures:

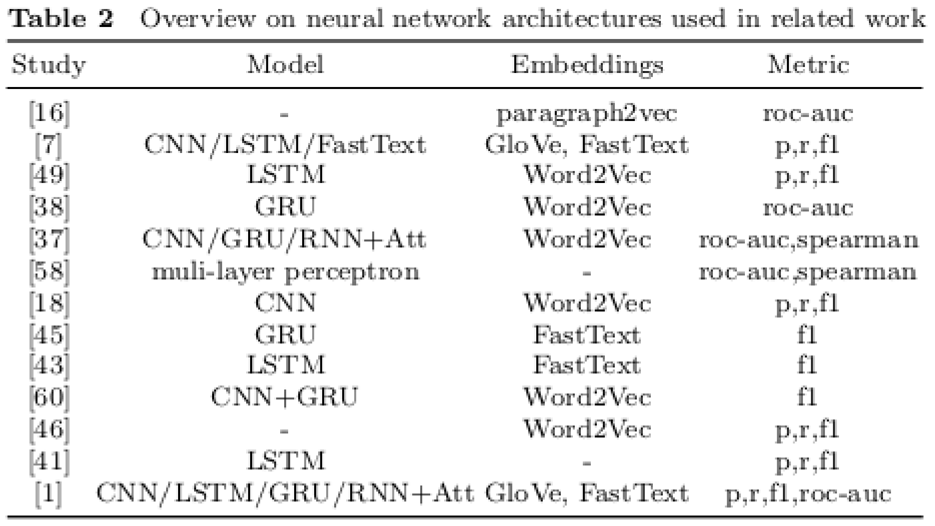


Table 2 Overview on neural network architectures used in related work

Study	Model	Embeddings	Metric
[16]	-	paragraph2vec	roc-auc
[7]	CNN/LSTM/FastText	GloVe, FastText	p,r,f1
[49]	LSTM	Word2Vec	p,r,f1
[38]	GRU	Word2Vec	roc-auc
[37]	CNN/GRU/RNN+Att	Word2Vec	roc-auc,spearman
[58]	muli-layer perceptron	-	roc-auc,spearman
[18]	CNN	Word2Vec	p,r,f1
[45]	GRU	FastText	f1
[43]	LSTM	FastText	f1
[60]	CNN+GRU	Word2Vec	f1
[46]	-	Word2Vec	p,r,f1
[41]	LSTM	-	p,r,f1
[1]	CNN/LSTM/GRU/RNN+Att	GloVe, FastText	p,r,f1,roc-auc

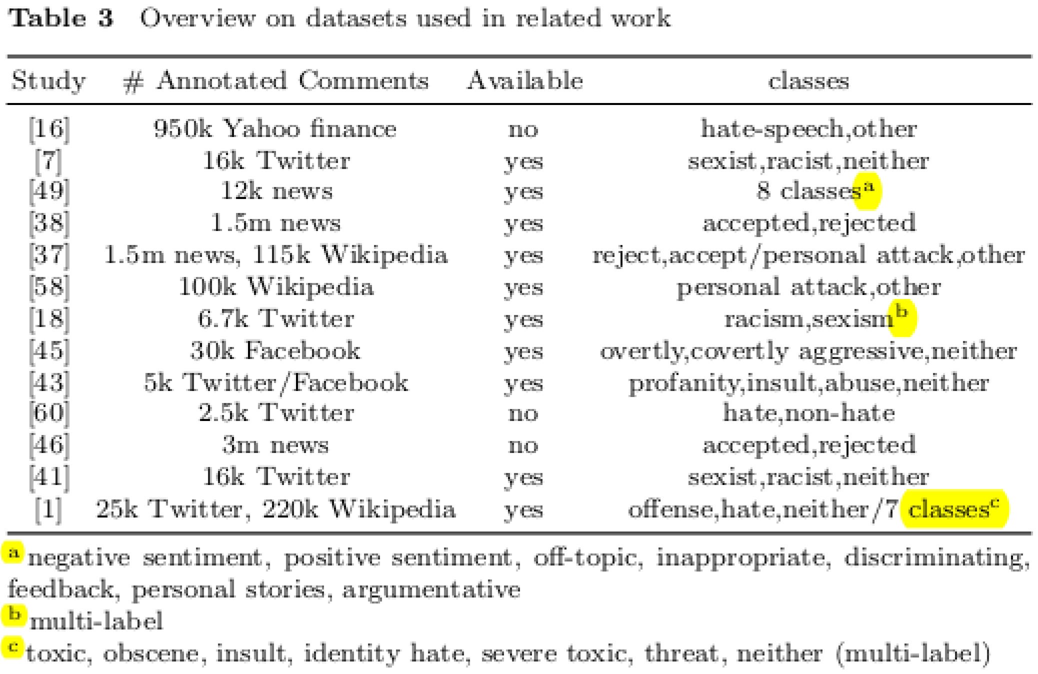


Table 3 Overview on datasets used in related work

Study	# Annotated Comments	Available	classes
[16]	950k Yahoo finance	no	hate-speech,other
[7]	16k Twitter	yes	sexist,racist,neither
[49]	12k news	yes	8 classes[a]
[38]	1.5m news	yes	accepted,rejected
[37]	1.5m news, 115k Wikipedia	yes	reject,accept/personal attack,other
[58]	100k Wikipedia	yes	personal attack,other
[18]	6.7k Twitter	yes	racism,sexism[b]
[45]	30k Facebook	yes	overtly,covertly aggressive,neither
[43]	5k Twitter/Facebook	yes	profanity,insult,abuse,neither
[60]	2.5k Twitter	no	hate,non-hate
[46]	3m news	no	accepted,rejected
[41]	16k Twitter	yes	sexist,racist,neither
[1]	25k Twitter, 220k Wikipedia	yes	offense,hate,neither/7 classes[c]

[a] negative sentiment, positive sentiment, off-topic, inappropriate, discriminating, feedback, personal stories, argumentative

[b] multi-label

[c] toxic, obscene, insult, identity hate, severe toxic, threat, neither (multi-label)

- **Spearman Correlation** is a rank correlation coefficient used to compare ground truth annotations with the model prediction. the correlation between the fraction of annotators voting in favor of toxic for a particular comment and the probability for the class toxic as predicted by the model is calculated.

From Binary to Fine-Grained Classification

- The question of `Does a particular comment need moderation or can it be published right away?` leads us to binary classification, but more fine-grained will give us why a comment is not suitable for publication.

- Therefore different classes of toxicity, such as insult, threat, obscene language, profane words, hate speech, etc. have to be distinguished.
- **Transfer Learning** Used to deal with limited training data and **explanations** used to help moderators to understand and trust neural network predictions.
- **Why it is hard problem?**:
    - language is not a statistic, and new words are introduced or words change their meaning
    - this flexible and ever-changing language requires a detection approach to adapt over time.
    - because of uncertainty researchers have come up with various annotation guidelines.
    - if we don't switch to more fine-grained labels 2 problems will appear;
        - Reduced available training data per class.
        - Increased difficulty for annotation.
    - the imbalance could cause bias in training NN, so `oversampling or downsampling or using weighted loss functions` to penalize for errors in minority class are made higher than the majority class. or another idea using `synthetic minority over-sampling technique(SMOTE)` . for `SMOTE and class weights similar gains in increased f1-score reported`
    - it is essential to keep trainable parameters and model capacity as small as possible if train data is limited. GRU 2 gates, LSTM 3 gates. so GRU is preferable because of small number of parameters. to aim to keep the number of parameters small explain the popularity of pooling layers. also, weight regularization can be used to limit the value range of parameters.
    - The inherent vagueness of language makes the annotation process even for domain experts, such as forum moderators, extremely difficult. Means the focus on training data generation lies on quality, not on quantity. on the other side, there is not much high quality annotated data available. one way to cope with these limitations is using transfer learning.
- **Transfer Learning**: the goal is to overcome low training data
    - The novel idea is to use machine-translation as an augmentation to convert English written text into another language and then revert it back and this could serve as another sample for training. (OBFUSCATION) . a dataset which obtained within this approach available online: https://hpi.de/naumann/projects/repeatability/text-mining.html
    - Another idea is to use transfer learning by fine-tuning models using pre-existed models and only fine-tune the latest layers using pre-existed data.
    - the paper `Attention Is All You Need` proposed an attention mechanism called a transformer.
    - with ELMo a technique to learn contextualized word embedding has been proposed
    - ULMFiT a fine-tuning method called `discriminative fine-tuning` has been introduced. which allows the transfer and applies pre-trained models to a variety of tasks.
    - BERT overcomes the limitation of all previous models that input needs to be processed sequentially left-to-right or right-to-left.
    - `More Importantly, The fine-grained classes can also provide a first explanation of why a comment is deleted` , It could delete specific comments due to some reasons(predicted classes) - It will increase trust in the machine-learned model.
- **EXPLANATIONS**:
    - > **this is what I was looking for!!!!!!!!!**
    - explanations are also very much needed to establish trust in the (semi-)automatic moderation process. Therefore, a fine-grained classification is inevitable.
    - it is necessary to point towards the phrases or words that make a comment off-topic, toxic, or insulting. These kinds of explanations are beneficial to monitor the algorithm and identify problems early on.

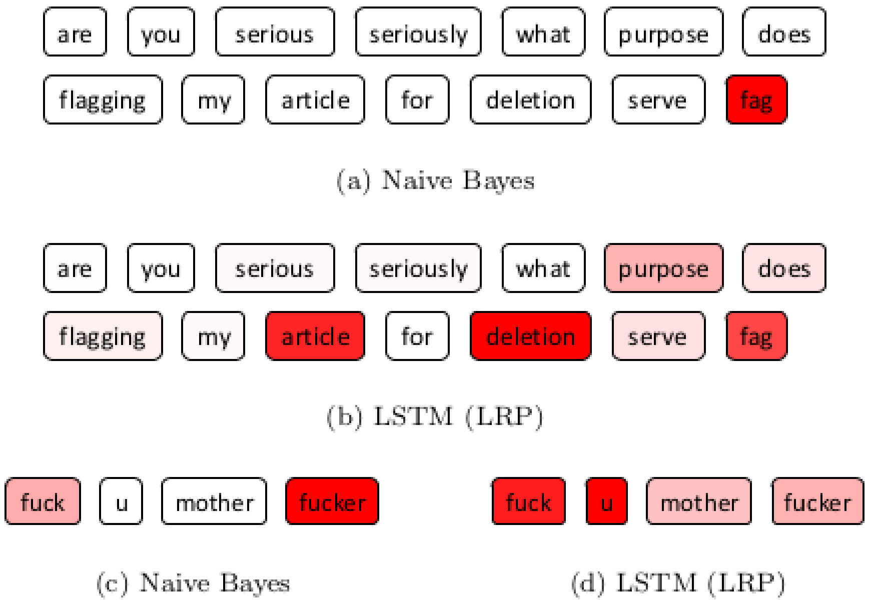


Fig. 1 Heatmaps highlight the most decisive words for the classification with a naive Bayes approach and an LSTM-based network.

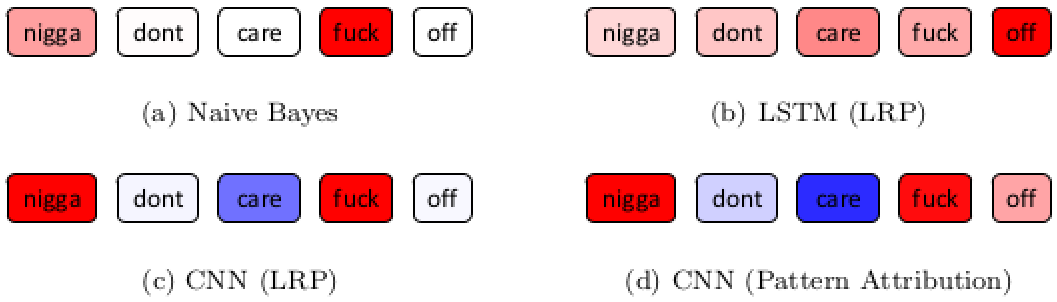


Fig. 2 Heatmaps highlight the most decisive words for the classification with a naive Bayes approach, an LSTM, and two CNNs.

> The visualizations are based on a tool called "innvestigate" by Alber et al.: https://github.com/albermax/innvestigate, red boxes indicate probability or relevance score in favor of the class "toxic", while blue boxes indicate the opposite class "not toxic"

- **Naive Bayes**: is a simple model that serves as a baseline approach for an explanation. for a few words in the vocabulary, we can calculate the probability that a comment with tat word is classified as a toxic. assumption of word independence is inherent to this approach. as result word correlations are not taken into account. so the probability for that word will be assigned across all comments.

- **Layer-wise Relevance Propagation (LRP)**: figure-1: The LRP visualization reveals that the LSTM correctly identifies word pairs that refer to each other.

Naive Bayes highlights only a small number of words as decisive for the classification. This problem is known as `over-localization` and has been reported as a problem also for other explanation approaches too.

Real-World Applications:

- Moderation is necessary to ensure respectful online discussions and to prevent misuse by spammers, haters, and trolls. And it is costly.

- **Semi-Automated Comment Moderation**: it can support human moderators but not completely replace them. the industry falls back to fewer complex models like logistic regression.
- **Troll Detection**: Trolls is `whose real intentions are to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement`
  - In contrast to toxic comment classification, the focus is on users who attract negative attention with multiple misbehavior.
  - malicious users, people with more accounts, people who try to increase the popularity of a post by its own fake accounts. or ... if I forget my username and make another account I will be considered as a malicious user.
  - DATASET: 3 million tweets by almost 3,000 Twitter troll accounts, Internet Research Agency (IRA) a Russian organization is characterized as a "troll factory" and is accused of having interfered with the U.S. presidential election in 2016 in a way that is prohibited by U.S. law.
  - Five different classes of IRA-associated Twitter accounts: Right Trolls(support Trump), Left Trolls(support Bernie Sanders), News Feeds(post local), Hashtag Gamers(post with specific hashtag), Fearmongers(spread of fear)

Current Limitations and Future Trends

- Common challenges are out-of-vocabulary words, long-range dependencies, and multi-word phrases[1]
- To cope with these challenges, sub-word embeddings, GRUs and LSTMs, and phrase mining techniques have been developed
- A detailed error analysis by van Aken reveals an open challenge.
- Examples of misclassification of comments from Wikipedia talk page and user comments

> (https://kaggle.com/c/jigsaw-toxic-comment-classification-challenge/):

- **Misclassification of Comments**: based on Van Aken[1] there are 6 common causes for misclassification:
  - False-positive is toxic comments that are misclassified as non-toxic
  - **Toxicity without swear words**: example: `she looks like a horse`, models need to understand that `she` refers to a person and `looking like a horse` is generally considered as an insulting. In contrast to these `False Negative`, there are `False positive`, example: `Oh, I feel like such an asshole now. Sorry, bud.`
  - **Questions, References, Metaphors, and comparisons**: models are not able to take the context of comments into account which includes other comments in the discussion. References comments could misclassified to toxic while they are referring to a comments that are toxic like: Example: `I deleted the Jews are dumb comment` (this is False Positive). example: `Who are you a sockpuppet for?` (False Negative). `sockpuppet` is not toxic in itself but addressing people as sockpuppet will make this example as a toxic.
  - **Sarcasm, Irony, and Rhetorical Questions**: Example comment: `hope you’re proud of yourself. Another milestone in idiocy.` the first sentences is nothing toxic, but second shows its opposite. rhetorical questions like `have you no brain?!?!` its insult. Rhetorical questions in toxic comments often contain subtle accusations, which current approaches hardly detect.
  - **Mislabeled Comments**: annotation of toxic comments is a challenging task because of different and diffcult guidelines to grasp every edge case. examples: `No matter how upset you may be there is never a reason to refer to another editor as ‘an idiot`, many state-of-the-art approaches classify this comment as not toxic. and it is labeled as a toxic. similar to this False Negative, there are False Positives too: `IF YOU LOOK THIS UP UR A DUMB RUSSIAN` which wrongly annotated as a non-toxic.
  - **Idiosyncratic and Rare Words**: obfuscated words, typos, slang, abbreviations, and neologisms are a particular challenge in toxic comment datasets. there are not enough samples from these words in the learned representations. example1: `fucc nicca yu pose to be pullin up` example2: `WTF man. Dan Whyte is Scottish` (depend on understanding `WTF`)

> misclassification due to rare words is twice as high for tweets than for Wikipedia talk page comments

Research Directions

- What is the opposite of toxic comments? High quality, engaging comments! Finding them automatically is a growing research field.
- Using context of comments - instead of using a single comments as input the full discussion and other context such as new article or user history can be used.
- Dealing with biased training data jigsaw competitions: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification
- The explainability of classification decisions. Good explanations are essential in semi automated comment moderation tools to help the moderators to make the right decision

CITE

```
@incollection{risch2020toxic,
  title={Toxic Comment Detection in Online Discussions},
  author={Risch, Julian and Krestel, Ralf},
  booktitle={Deep Learning-Based Approaches for Sentiment Analysis},
  pages={85--109},
  year={2020},
  publisher={Springer}
}
```