SemEval-2021 Task 5: Toxic Spans Detection

John Pavlopoulos[†⋆], **Jeffrey Sorensen**[‡]
Léo Laugier[◇], **Ion Androutsopoulos**[†]
[⋆] Department of Computer and System Sciences, Stockholm University, Sweden
[†] Department of Informatics, Athens University of Economic and Business, Greece
annis,ion@aueb.gr
[◇] Télécom Paris, Institut Polytechnique de Paris, France
leo.laugier@telecom-paris.fr
[‡] Google Jigsaw
sorenj@google.com

Abstract

The Toxic Spans Detection task of SemEval-2021 required participants to predict the spans of toxic posts that were responsible for the toxic label of the posts. The task could be addressed as supervised sequence labeling, using training data with gold toxic spans provided by the organisers. It could also be treated as rationale extraction, using classifiers trained on potentially larger external datasets of posts manually annotated as toxic or not, without toxic span annotations. For the supervised sequence labeling approach and evaluation purposes, posts previously labeled as toxic were crowd-annotated for toxic spans. Participants submitted their predicted spans for a held-out test set, and were scored using character-based F1. This overview summarises the work of the 36 teams that provided system descriptions.

1 Introduction

Discussions online often host toxic posts, meaning posts that are rude, disrespectful, or unreasonable; and which can make users want to leave the conversation (Borkan et al., 2019a). Current toxicity detection systems classify whole posts as toxic or not (Schmidt and Wiegand, 2017; Pavlopoulos et al., 2017; Zampieri et al., 2019), often to assist human moderators, who may be required to review only posts classified as toxic, when reviewing all posts is infeasible. In such cases, human moderators could be assisted even more by automatically highlighting spans of the posts that made the system classify the posts as toxic. This would allow the moderators to more quickly identify objectionable parts of the posts, especially in long posts, and more easily approve or reject the decisions of the toxicity detection systems. As a first step along this direction, Task 5 of SemEval 2021 provided the participants with posts previously rated to be toxic, and required them to identify toxic spans, i.e., *spans that were responsible for the toxicity of the posts*, when identifying such spans was possible. Note that a post may include no toxic span and still be marked as toxic. On the other hand, a non toxic post may comprise spans that are considered toxic in other toxic posts. We provided a dataset of English posts with gold annotations of toxic spans, and evaluated participating systems on a held-out test subset using character-based F1. The task could be addressed as supervised sequence labeling, training on the provided posts with gold toxic spans. It could also be treated as rationale extraction (Li et al., 2016; Ribeiro et al., 2016), using classifiers trained on larger external datasets of posts manually annotated as toxic or not, without toxic span annotations. There were almost 500 individual participants, and 36 out of the 92 teams that were formed submitted reports and results that we survey here. Most teams adopted the supervised sequence labeling approach. Hence, there is still scope for further work on the rationale extraction approach. We also discuss other possible improvements in the definition and data of the task.

2 Competition Dataset Creation

During 2015, when many publications were closing down comment sections due to moderation burdens, a start up named Civil Comments launched (Finley, 2016). Using a system of peer-based review and flagging, they hoped to crowd source the moderation responsibility. When this effort shut down in 2017 (Bogdanoff, 2017), they cited the financial constraints of the competitive publishing industry and the challenges of attaining the necessary scale.

The founders of Civil Comments, in collaboration with researchers from Google Jigsaw, undertook an effort to open source the collection of more than two million comments that had been collected. After filtering the comments to remove personally

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 59–69
Bangkok, Thailand (online), August 5–6, 2021.

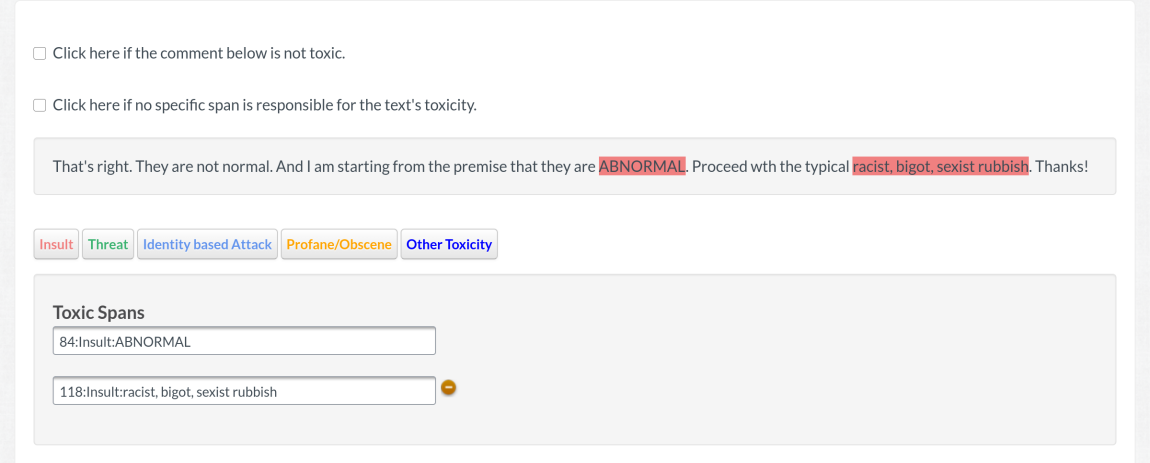

Figure 1: Screenshot of the Appen labeling interface that was used to annotate toxic spans.

identifiable information, a revised version of the annotation system of Wulczyn et al. (2017) was used on the Appen crowd rating platform to label the comments using a number of attributes including 'toxicity', 'obscene', 'threat' Borkan et al. (2019a). The complete dataset, partitioned into training, development, and test sets, was featured in a Kaggle competition,[1] with additional material, including individual rater decisions, published (Borkan et al., 2019b) after the close of the competition.

Civil Comments contains about 30k comments marked as toxic by a majority of at least three crowd raters. Toxic comments are *rare*, especially in fora that are not anonymous and where people have expectations that moderators will be watching and taking action. We undertook an effort to re-annotate this subset of comments at the span level, using the following instructions:

> *For this task you will be viewing comments that a majority of annotators have already judged as toxic. We would like to know what parts of the comments are responsible for this.*
>
> *Extract the toxic word sequences (spans) of the comment below, by highlighting each such span and then clicking the right button. If the comment is not toxic or if the whole comment should have been annotated, check the appropriate box and do not highlight any span.*

and a custom JavaScript based template,[2] which allowed selection and tagging of comment spans (Fig. 1). While raters were asked to categorize each span as one of five different categories, this was primarily intended as a priming exercise and all of the highlighted spans were collapsed into a single category. The lengths of the highlighted spans were decided by the raters. Seven raters were employed per post, but there were posts where fewer were eventually assigned. On the test subset (Table 1), we verified that the number of raters per post varied from three to seven; on the trial and train subsets this number varied from two to seven. All raters were warned the content might by explicit, and only raters who allowed adult content were selected.[3]

2.1 Inter-annotator Agreement

We measured inter-annotator agreement, initially, on a small set of 35 posts and we found 0.61 average Cohen's Kappa. That is, we computed the mean pairwise Kappa per post, by using character offsets as instances being classified in two classes, toxic and non-toxic. And then we averaged Kappa over the 35 posts. On later experiments with larger samples (up to 1,000 posts) we observed equally moderate agreement and always higher than 0.55. Given the highly subjective nature of the task we consider this agreement to be reasonably high.

2.2 Extracting the ground truth

Each post comprises sets of annotated spans, one per rater. Each span is assigned a binary (toxic, non-toxic) label, based on whether the respective rater

[1] www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[2] github.com/ipavlopoulos/toxic_spans

[3] The full dataset and annotations for ToxicSpans is released (github.com/ipavlopoulos/toxic_spans) with a CC0 licence. The previously released Civil Comments dataset, on which the new dataset is based, was filtered to remove any potential personally identifiable information.

	Trial	Train	Test
Number of posts	690	7,939	2,000
Avg. post length	199.47	204.57	186.41
Avg. toxic span length	10.78	13.11	7.89
Avg. # of toxic spans	1.43	1.39	0.92

Table 1: Statistics of the trial, training, and test subsets of the dataset. Lengths are calculated in characters.

found the span to be insulting, threatening, identity-based attack, profane/obscene, or otherwise toxic. If the span was annotated with any of those types, the span is considered toxic according to the rater, otherwise not. For each post, we extracted the character offsets of each toxic span of each rater. In each post, the ground truth considers a character offset as toxic if the majority of the raters included it in their toxic spans, otherwise the ground truth of the character offset is non-toxic. A toxic span (Table 1) in the ground truth of a post is a maximal sequence of contiguous toxic character-offsets.

2.3 Exploratory analysis

After discarding duplicates and posts used as quiz questions to check the reliability of candidate annotators, we split the data into trial, train, and test (Table 1). Compared to the trial and training sets, the test set comprises posts with fewer characters and spans, but also shorter spans on average.

When studying the toxicity subtypes, we find that the vast majority of posts are annotated as insulting. In the training set, more than 6,000 posts are annotated as insulting, and the same high fraction is observed in the trial and test sets. Most of the toxic spans in the training set are single-word terms. The most frequent of them, such as 'stupid' and 'idiot', occur hundreds of times and remain frequent in the trial and test sets. Multi-word terms, such as 'white trash', 'mentally ill', are less frequent and vary across the three sets.

In an analysis of the test set, Palomino et al. (2021) used an emotion classifier that returns five scores per post, one for each of the following emotions: *anger*, *happiness*, *sadness*, *surprise*, *fear*.[4] *Fear* and *sadness* were reported to be the emotions with the highest average scores, a finding that we verified by repeating the experiment (see Fig. 2).[5] Interestingly, the emotion with the highest average score after *sadness* and *fear* is *surprise*, not *anger*, and *happiness* has the lowest score.

[4] pypi.org/project/text2emotion

[5] A post with a high *sadness* score (100%) is the following: "Such thin skin. Pathetic."; the toxic span shown in red.

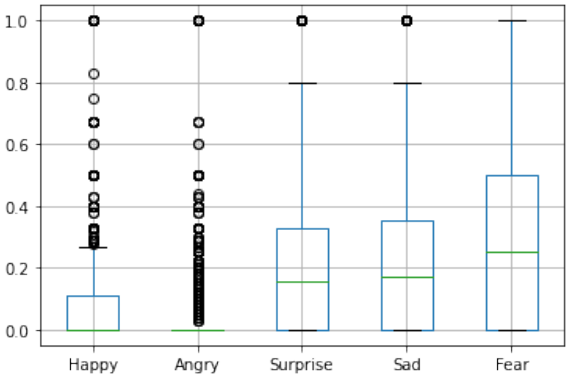

Figure 2: Emotion scores of the test posts. Emotion scores were obtained using an off-the-shelf emotion classifier, following Palomino et al. (2021).

3 Task description

The objective of this task is the detection of the spans that make a post toxic, when detecting such spans is possible. Systems had to extract a list of toxic spans, or an empty list, per post. A toxic span was defined to be a sequence of words that attribute to the post's toxicity. Although we defined the task at the word level, gold labels were provided at the character level counting from zero (see Table 2).

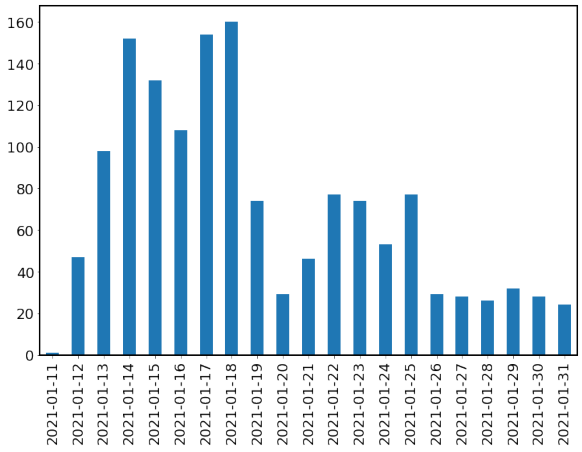

Figure 3: Number of submissions per evaluation day.

The evaluation period started on January 10, 2021 and finished on January 31, 2021. The first week, 10 submissions were allowed per day per team. The second week this number was reduced to 5, and it was reduced further to 1 during the final week. We chose to allow an extended evaluation period combined with multiple team submissions, to promote the competition. However, we also chose a decreasing submission limit to make it harder for participants to overfit the test set. As shown in Fig. 3, the number of submissions dropped over time, due to this constraint, but the interest was

OFFSETS	POST
{}	I will stand by my statement that liberalism is what smart looks like to stupid people. The letter posted yesterday states the real problem clearly enough. "It doesn't matter if it was protected by the First Amendment." I have legal news flash for the University of Oregon. Because the University is public and receives government tax dollars it is indeed bound to abide by the US constitution with All OF IT'S PROTECTION INCLUDED. Which should tell everyone right off the signatories are anti-American, and perhaps even anti-America in their twisted PC nonsense.
{}	Fits the despicable piece of trash like a glove.
{}	No he shot someone so he shouldn't be free. He should be locked up like He is for the next decade. You holla like its no big deal but what if it was your little brother who was shot? Then your dumb ass wouldnrlt be screaming free ace g or whatever gay nickname he has. That is the lamest gangster name ever. Do you think that name scares people? Ace clown is what i will now call him.
{55, 56, 57, 58}	Time to give the opposition weapons and overthrow this jerk.
{31, . . . , 39}	you are a paid liberal hack of stupidity shore living the liberal lie every day
{9, . . . , 13}	Youre an idiot.
{7, 8, . . . , 11, 14, 15, . . . , 18}	what a dummy, dummy has no regard for the law. you cannot brutalize a suspect. he has complete lack of respect for any law and is acting like a dictator. he is trying to emulate putin.
{12, . . . , 17, 94, . . . , 102}	People make stupid decisions and then expect the gov't to bail them out. There is no cure for stupidity.
{14, . . . , 20, 29, . . . , 35}	Nah, the only asshole is the asshole firing a rifle within city limits.

Table 2: Examples of toxic test posts and their ground truth toxic spans (shown in red). The left column shows the character offsets of the toxic spans. The top three posts have no toxic spans, the next three have one each, while the remaining three posts have two toxic spans each.

continuous, and there were submissions until the last day. Despite the decreasing total number of submissions per day, the top daily score increased, reaching its maximum on the last day (see Fig. 4).

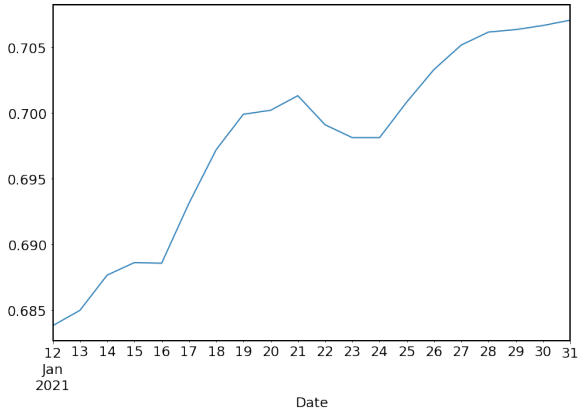

Figure 4: The evaluation score (character F1) of the best submission per day during the evaluation period.

4 Participation overview

We received 479 individual participation requests, 92 team formations, and 1,449 submissions. 91 teams submitted valid predictions (1,385 valid submissions in total) and were scored; out of these, only 36 submitted system descriptions.

4.1 The HITSZ-HLT submission

The best performing team (HITSZ-HLT) formulated the problem as a combination of token labeling and span extraction (Zhu et al., 2021).

For their token labeling approach, the team used two systems based on BERT (Devlin et al., 2019). Both systems had a Conditional Random Field (CRF) layer (Sutton and McCallum, 2006) on top, but one of the two also had an LSTM layer (Hochreiter and Schmidhuber, 1997) between BERT and the CRF layer. In both approaches, word-level BIO tags were used, i.e., words were labelled as B (beginning word of a toxic span), I (inside word of a toxic span), or O (outside of any toxic span).

For their span extraction approach, the team also used BERT. Roughly speaking, in this case BERT produces probabilities indicating how likely it is for each token to be the beginning or end of a toxic span. Then a heuristic search algorithm, originally developed for target extraction in sentiment analysis by Hu et al. (2019), selects the best combinations of candidate begin and end tokens, aiming to output the most likely set of toxic spans per post.

The character predictions of the three systems described above were combined with majority voting per character. That is, if any two systems considered a character to be part of a toxic span, then the ensemble classified the character as toxic, otherwise the ensemble classified it as non-toxic.

4.2 The S-NLP submission

The team with the second best performing system (S-NLP) consists of individual participants who grouped and submitted an ensemble of their sys-

tems (Nguyen et al., 2021). The ensemble combines two approaches, both of which are based on a RoBERTa model (Liu et al., 2019). The latter is first fine-tuned to classify posts as toxic or non-toxic, using three Kaggle toxicity datasets.[6] For toxic span detection, RoBERTa's subword representations from three different layers (1, 6, 12) are summed to produce the corresponding word embeddings. A binary classifier on top of RoBERTa, operating on the word embeddings, predicts whether a word belongs to a toxic span or not.

For the first component of the ensemble, the word embeddings obtained from RoBERTa's subword representations are concatenated with FLAIR (Akbik et al., 2019) and FastText (Bojanowski et al., 2017) embeddings.[7] The resulting embeddings are passed on to a two-layer stacked BiLSTM with a CRF layer on top to generate a BIO tag per word.

The second component of the ensemble used the RoBERTa model as a teacher to produce silver toxic spans for 30,000 unlabelled toxic posts (Borkan et al., 2019a). RoBERTa was then retrained as a student on the augmented dataset (30k posts with silver labels and the training posts provided by the organisers) to predict toxic offsets.

The ensemble returns the intersection of the toxic spans identified by the two components.

4.3 Additional interesting approaches

We now discuss some of the most interesting alternative approaches tried by the participants, even if they did not lead to high scores.

Rationales Some participants experimented with training toxicity classifiers on external datasets containing posts labeled as toxic or non-toxic; and then employing model-specific or model-agnostic rationale extraction mechanisms to produce toxic spans as explanations of the decisions of the classifier. The model-specific rationale mechanism of Rusert (2021) used the attention scores of an LSTM toxicity classifier to detect the toxic spans. Pluciński and Klimczak (2021) used the same approach, but also employed an orthogonalisation technique (Mohankumar et al., 2020). The model-agnostic rationale mechanism of Rusert (2021) combined an LSTM classifier with a token-masking approach that we call Input Erasure (IE), due to its similarities to the method of Li et al. (2016). The model-agnostic approach of Pluciński and Klimczak (2021) combined SHAP (Lundberg and Lee, 2017) with a fine-tuned BERT model. Ding and Jurgens (2021) and Benlahbib et al. (2021) also experimented with model-agnostic approaches, but they combined LIME (Ribeiro et al., 2016) with a Logistic Regression (LR) or with a linear Support Vector Machine (SVM) toxicity classifier. All the above mentioned approaches used a threshold to turn the explanation scores (e.g., attention or LIME scores) of the words into binary decisions (toxic/non-toxic words).

Lexicon-based No team relied on a purely lexicon-based approach, but few experimented with lexicon-based baselines (Zhu et al., 2021; Palomino et al., 2021) or used such components in ensembles (Ranasinghe et al., 2021). Three kinds of lexicon-based methods were used. First, the lexicon was handcrafted by domain experts (Smedt et al., 2020) and it was simply employed as a list of toxic words for lookup operations (Palomino et al., 2021). Second, the lexicon was compiled using the set of tokens labeled as toxic in our span-annotated training set and it was used as a lookup table (Burtenshaw and Kestemont, 2021), possibly also storing the frequency of each lexicon token in the training set (Zhu et al., 2021). The former two were also combined (Ranasinghe et al., 2021). Third, the least supervised lexicons were built with statistical analysis on the occurrences of tokens in a training set solely annotated at the comment level (toxic/non-toxic post) (Rusert, 2021). An added value of these approaches is that easy to use resources (toxicity lexicons) are built and shared publicly, such as the one suggested by Pluciński and Klimczak (2021).[8]

Custom losses Zhen Wang and Liu (2021) experimented with a new custom loss, which weighted false toxicity predictions based on their location in the text. If a false prediction was located near a ground truth toxic span, then it would contribute less to the overall loss for that post, compared to one located further away. The loss function used by Kuyumcu et al. (2021) to train their system is the Tversky Similarity Index (Tversky, 1977), a generalisation of the Sørensen–Dice coefficient and the Jaccard index, which was adjusted by the authors to weigh up false negatives.

Data augmentation The vast majority of the participating teams employed additional training data annotated at the post level. That is, either to

[6] github.com/unitaryai/detoxify

[7] In the latter case, in-vocabulary word embeddings were imported to Word2Vec for efficiency, and out of vocabulary words were handled with BPEs (Sennrich et al., 2016).

[8] github.com/Orthrus-Lexicon/Toxic

build lexicons (Rusert, 2021), to leverage unsupervised rationale extraction methods (Rusert, 2021; Pluciński and Klimczak, 2021; Ding and Jurgens, 2021; Benlahbib et al., 2021), or to filter posts (Luu and Nguyen, 2021) that were not labeled as toxic by a toxicity classifier. Suman and Jain (2021) astutely produced silver data from external sources to augment the initial golden annotated dataset, training their model iteratively in a semi-supervised manner.

5 Evaluation

This section focuses on the evaluation framework of the task. First, the official measure that was used to evaluate the participating systems is described. Then, we discuss baseline models that were selected as benchmarks for comparison reasons. Finally, the results are presented.

5.1 Official evaluation measure

Following the work of Martino et al. (2019), systems were evaluated in terms of F1 computed on character offsets. For each system, we computed the F1 score per post, between the predicted and the ground truth character offsets. Then, we returned the macro-averaged (over test posts) score. When the ground truth set of character offsets was empty, we assigned a perfect score ($F_1 = 1$) to the post in question if the predicted set of character offsets was also empty, and a zero score otherwise.[9]

5.2 Benchmarks

We report the results of some baselines, developed by us or the participants, to act as benchmarks.

BENCHMARK I was developed by Nguyen et al. (2021). It is based on a RoBERTa model, fine-tuned to predict if a post is toxic or not (Section 4.2) and further fine-tuned to predict toxic spans by using a CRF layer on top.

BENCHMARK II is a lexicon-based system, developed by Zhu et al. (2021), which extracts likely toxic words from the training data and simply tags them during inference. The lexicon comprises words that appear frequently inside ground truth toxic spans and not outside.

BENCHMARK III is a random baseline, which assigns a random label (toxic/non-toxic) per character offset (50% chance of being toxic).[10]

[9]The evaluation code can be found in our GitHub repository (`github.com/ipavlopoulos/toxic_spans`).

[10]The code of this baseline is also in the task's repository.

5.3 Results

RANK	TEAM	SCORE (%)
1	HITSZ-HLT	70.83
2	S-NLP	70.77
BASELINE	BENCHMARK I	69.89
3	hitmi&t	69.85
5	YNU-HPCC	69.63
7	Cisco	69.22
8	MedAI	69.03
9	IITKDetox	68.95
13	GHOST	68.59
14	HLE-UPC	68.54
15	UTNLP	68.44
16	YoungSheldon	68.42
17	Lone Pine	68.38
18	sk	68.32
20	WLV-RIT	68.01
21	CSECUDSG	67.95
22	LISAC FSDM USMBA	67.84
23	UoT-UWF-PartAI	67.70
25	uob	67.61
MEDIAN	The median score	67.58
26	UAntwerp	67.55
27	MIPT-NSU-UTMN	67.55
28	NLRG	67.53
30	HamiltonDinggg	67.15
33	lz1904	67.00
34	UIT-E10dot3	66.99
36	UniParma	66.72
37	hub	66.40
38	GoldenWindPlymouth	66.37
41	AStarTwice	66.16
44	sefamerve_arge	66.01
46	UPB	65.73
49	Entity	65.61
BASELINE	BENCHMARK II	64.98
57	BennettNLP (Fuchsia)	64.53
58	TeamGriek	64.31
63	UIT-ISE-NLP	62.23
75	NLP_UIowa	50.09
BASELINE	BENCHMARK III	12.22
90	macech	7.33

Table 3: Official rank and F1 score (%) of the 36 participating teams that submitted system description papers. (There were 91 teams with sumbissions in total.) The median is shown in blue and benchmarks in red.

Table 3 shows the scores and ranks of all participating teams that described their approach, i.e., 36 out of 91 teams that participated. HITSZ-HLT (Section 4.1) was ranked first, followed by S-NLP (Section 4.2) that scored 0.06% lower. The rest of the teams followed with scores lower than 70%.

The score of the median is 67.58%, which is not far below the top scored team (-3.22 percent units), while it is far above the last two (+17.52 percent units). The standard deviation of system scores above the median is much lower (0.94) than that of the systems below the median (4.12). Most teams that were excluded from the table (because they did not describe their methods) score lower than

the median. However, there were also top scoring teams among those that were excluded, such as a team with a RoBERTa-based token-level ensemble that was ranked 4th.[11]

BENCHMARK I achieves a considerably high score and, hence, is very highly ranked. Combining BERT with a CRF or a span extraction method (two of the individual methods of the HITSZ-HLT ensemble, Section 4.1, not shown in Table 3) also performs well (Zhu et al., 2021), but these methods would be ranked two positions lower than BENCHMARK I. Nguyen et al. (2021) explored the benefits of further enhancing these word embeddings by concatenating them with FLAIR (Akbik et al., 2019) and FastText (Bojanowski et al., 2017) embeddings (Section 4.2). As shown in Fig. 5, the F1 score is slightly improved, reaching a maximum when both FLAIR and FastText embeddings are added.[12] We note that the same beneficial effect of enhancing the word embeddings was reported when using BERT as the base model (Sans and Farràs, 2021).

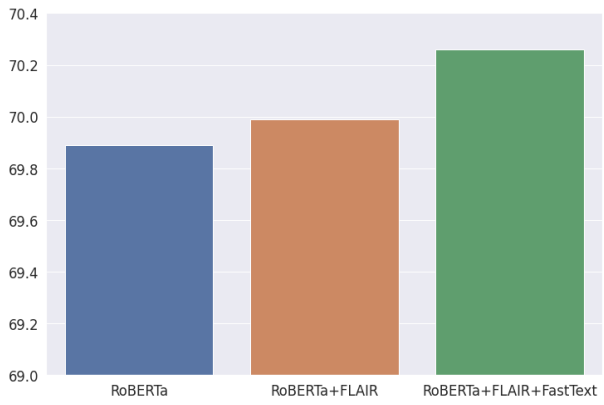

Figure 5: F1 of BENCHMARK I (Zhu et al., 2021) when FLAIR and FastText word embeddings are concatenated with the embeddings obtained from RoBERTa's subword representations (from layers 1, 6, 12).

The lexicon-based BENCHMARK II and the random BENCHMARK III scored very low. The latter outperformed only one submission (MACECH), which sent the predictions in the wrong order. As noted in their report (Cech, 2021), if the predictions had been submitted in the correct order, the team's score would have been 54%, and BENCHMARK III would have been the worst system in Table 3.

[11] We asked for details from participants that did not submit a description paper, but not all of them replied.

[12] Out of vocabulary words were tackled by using FastText embeddings of BPEs; consult Nguyen et al. (2021).

6 Analysis and discussion

Overall the organisers were happy to see the degree of involvement in this shared task, and the resulting diversity of approaches to this problem. We include some of our observations regarding the administration of the evaluation and what we have learned from the results.

6.1 Participation

The authors reached out to teams that decided not to submit a description paper and the vast majority were students who were time-limited. The fact that students participated in the task is promising and we plan to consider more ways to introduce SemEval tasks in classrooms. On the other hand, 60% of the participants chose not to describe their approach, which is problematic and should be addressed. A team could take advantage of such an option to create duplicate submissions and bypass any submission limits. More importantly, potentially interesting approaches are not discussed and properly compared to others.

It is also worth mentioning that the extended timeline allowed participants to join forces. For instance, a number of participants decided to combine their systems and form the 2nd ranked S-NLP. Their ensemble scored higher than all their standalone systems, though their best standalone system would still be ranked 2nd. In any case, we welcome the collaboration between participants, which may provide further insights regarding effective combinations of architectures.

6.2 General remarks on the approaches

Except for lexicon-based baselines, we observed that the vast majority of systems adopted the recent paradigm in NLP: fine-tuning large off-the-shelf Transformers (Vaswani et al., 2017) pre-trained on massive corpora. Non-Transformer based approaches, mostly LSTMs with pre-trained word embeddings were also used. The nature of the task, similar to the well-studied Named Entity Recognition (NER) task, led many competitors to use a CRF layer on top of the model (e.g., Transformers or LSTMs) of their choice.

6.3 Performance

The winning team (HITSZ-HLT) combined BERT with two approaches for their ensemble: a token labeling approach (two versions, with/without an LSTM between BERT and the CRF) and a span ex-

traction approach (Section 4.1). The comparison of the two showed that span extraction is slightly better on posts with a single span, but token labeling is clearly better on multi-span posts (Zhu et al., 2021). The complementary nature of the two approaches is probably what makes even a simple majority voting ensemble better than its competitors.

The system that was ranked second (S-NLP) also employed an ensemble, using a RoBERTa model initially fine-tuned to classify posts as toxic or non-toxic as the starting point (Nguyen et al., 2021). The ensemble combined (i) the resulting RoBERTa model, now fine-tuned to predict toxic spans, with additional FLAIR and FastText embeddings, and (ii) a RoBERTa model retrained as a student to predict toxic spans (Section 4.2). Although the two standalone models achieved higher scores than the standalone models of the top-ranked team (HITSZ-HLT), the ensemble did not yield significant improvements. This may be due to the student's decisions not being that complementary to the teacher's, as the team notes (Nguyen et al., 2021).

TBC	RE	F1 (%)	Report
LSTM	IE	38.29	Rusert (2021)
LSTM	ATT	49.70	Pluciński and Klimczak (2021)
LSTM	ATT	50.07	Rusert (2021)
LR	LIME	58.88	Benlahbib et al. (2021)
SVM	LIME	59.21	Benlahbib et al. (2021)
BERT	SHAP	**59.87**	Pluciński and Klimczak (2021)

Table 4: F1 on the evaluation set for systems employing rationale extraction (RE) mechanisms combined with post-level toxicity binary classifiers (TBC). Rationales are obtained via Input Erasure (IE), Attention (ATT), LIME, or SHAP. The binary classifier is an LSTM, Logistic Regression (LR), SVM, or BERT.

Teams that experimented with rationale extraction mechanisms (Section 4.3) did not find this approach advantageous compared to supervised sequence labeling in terms of F1 scores. However, the reported results of the rationale-based systems show that this approach is promising, especially because it does not require any data annotated at the span-level. Hence, there is scope for future work that could explore this direction further. Table 4 shows the F1 scores of all the rationale-based systems that were reported by participants. The binary toxic post classifiers that were used were LSTM, Logistic Regression (LR), Support Vector Machines (SVM), and BERT. The attention scores of an LSTM were used with (Pluciński and Klimczak, 2021) and without an orthogonality method (Rusert, 2021), with the latter being slightly better; these are model-specific rational extraction methods (Section 4.3). Model-agnostic approaches (Input Erasure, LIME, SHAP) were better than the model-specific ones. The best rationale-based method employed a BERT model, fine-tuned for toxic post classification, and SHAP.

Lexicon Name	F1 (%)	Report
WIEGAND 1 †	33.07	Zhu et al. (2021)
WORD-MATCH	40.86	Ranasinghe et al. (2021)
FREQ-RATIO †	41.55	Rusert (2021)
LOOKUP ‡	41.61	Burtenshaw and Kestemont (2021)
WIEGAND 2 †	50.98	Zhu et al. (2021)
ORTHRUS	61.07	Palomino et al. (2021)
HITSZ-HLT ‡	**64.98**	Zhu et al. (2021)
+WORDNET	64.09	Zhu et al. (2021)
+GLOVE	64.19	Zhu et al. (2021)

Table 5: F1 on the evaluation set for lexicon-based systems. Systems that are followed by † and ‡ use exclusively external and internal resources respectively.

Lexicon-based approaches were only used as baselines or components in ensembles, as already noted. In principle, all lexicon-based systems are extremely efficient and interpretable. Table 5 shows they can also achieve surprisingly high scores. Recall that we used the best performing lexicon-based system, developed by Zhu et al. (2021), as BENCHMARK II. Its score is included in Table 3. Despite the fact that it is low ranked, its F1 score is less than 6 percent points lower that that of the best submission. We also note that BENCHMARK II is a high-precision classifier; it outperforms even the best system in terms of precision (Zhu et al., 2021). Attempts to expand its lexicon using WordNet and GloVe, improved recall, but eventually harmed precision and its F1 score.

6.4 Error analysis

A common theme across many competitor reports was the serious challenge posed by comments with no toxic spans. It is not readily evident why this is a common occurrence in the task, and certainly the way that annotation consensus is used to combine annotations can be a contributing factor. However, many systems seemed determined to tag *some* spans and many authors noted that performance on posts with no tagged span was extremely poor compared to performance on posts with tagged spans.

Many systems were also reluctant to tag function words like 'of' and 'and', which can be included in multi-word spans (e.g., 'piece of crap'), leading to a decline in performance as measured by the chosen F1 measure. The overwhelming presence

of single word gold spans in the training set favors short spans. But the majority of the short spans comprises common cuss or clearly abusive words, which can be directly classified as toxic (Ghosh and Kumar, 2021); by contrast, the infrequent longer spans are rather context dependent and more challenging to detect. This probably also contributed to the performance of the best system (HITSZ-HLT), since one of the two components of that ensemble handled better long spans, as already discussed in Section 6.3.

Other error analysis highlighted challenges intrinsic to the task. The strong dependency of toxicity on context makes it particularly difficult to solve with systems based on vocabulary. Toxicity, when expressed with subtle language, can appear through non-local text features: some comments are toxic without showing any obvious toxic span in them. Such posts made the task more difficult for participants, because systems had learnt to label the words bearing the most negative sentiment (Bansal et al., 2021). Annotation mistakes were also reported (Table 6).

Type	Description
INCONSISTENCIES	Not all the occurrences of the same toxic span are annotated in the same post.
FALSE NEGATIVES	Toxic words missed.
FALSE POSITIVES	Non-toxic words labelled.

Table 6: The types and descriptions of the annotation mistakes that were detected by some of the participants.

Participants that were notable for their effort in error analysis include Bansal et al. (2021), Hoang and Nguyen (2021), Ding and Jurgens (2021), and Ghosh and Kumar (2021), where an additional effort was made to examine their model's ability to correctly tag words in toxic and non-toxic contexts. Interestingly Sans and Farràs (2021) also noted in their analysis that racial and ethnic terms are labeled in biased ways that reflect patterns not only in the training toxic spans, but also in external data used to pre-train underlying Transformer models.

7 Conclusions

We provided 10,629 posts that were annotated for toxic spans and we defined the task of toxic span detection. The task was popular, attracting almost 500 individual participants. Eventually 91 teams were formed, out of which 36 submitted a description report. This overview described the approaches of these 36 teams and discussed their results.

Pre-trained Transformers, fine-tuned by viewing the task as a sequence labelling one, performed well and solutions that combined these models within an ensemble were highly-rated. The performance of these models increases further with the help of pre-trained word embeddings or by using multiple Transformer layers to embed words.

Long toxic spans were more likely context-dependent and less frequent in the dataset compared to single-word spans, which made their detection a challenge. The winners included in their ensemble an approach that performed better on long spans, but we note that the problem of detecting long uncommon toxic spans is far from solved.

Of particular interest were approaches that employed rationale extraction mechanisms, which do not require any training data annotated at the span level. They performed much worse than sequence labeling approaches, but this is a promising direction that was considered by only a few participants.

Future similar competitions could benefit from tracks that separate supervised from unsupervised solutions. The development of datasets created with the help of crowd annotators should focus on addressing ambiguity, bias, inconsistencies, and misannotations. This could be accomplished by adding more annotators per post. Future competitions could also require participants to both classify posts as toxic or not, and detect toxic spans only when posts are classified as toxic, instead of providing the participants only with posts already classified as toxic. Finally, future competitions could require participants to distinguish toxic posts of different kinds (e.g., insult, threat, profanity, along with supporting spans), which are sometimes easier to define compared to the more general umbrella toxicity term we (and others) have used.

Acknowledgement

We thank the participants and the reviewers for their useful comments and suggestions. This research was funded in part by a Google Research Award.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL Demonstrations*, pages 54–59.

Archit Bansal, Abhay Kaushik, and Ashutosh Modi. 2021. IITK@Detox at SemEval-2021 Task 5: Semi-

supervised learning and dice loss for toxic spans detection. In *SemEval*.

Abdessamad Benlahbib, Hamza Alami, and Ahmed Alami. 2021. LISAC FSDM USMBA at SemEval 2021 Task 5: Tackling toxic spans detection challenge with supervised spanBERT-based model and unsupervised LIME-based model. In *SemEval*.

Aja Bogdanoff. 2017. Saying goodbye to civil comments. Accessed: 2021-04-15.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019a. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW*, pages 491–500, San Francisco, USA.

Daniel Borkan, Jeffrey Sorensen, and Lucy Vasserman. 2019b. Exploring the role of human raters in creating nlp datasets. Accessed: 2021-04-15.

Ben Burtenshaw and Mike Kestemont. 2021. UAntwerp at SemEval-2021 Task 5: Spans are spans, stacking a binary word level approach to toxic span detection. In *SemEval*.

Maggie Cech. 2021. macech at SemEval-2021 Task 5: Toxic spans detection. In *SemEval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Huiyang Ding and David Jurgens. 2021. HamiltonDinggg at SemEval-2021 Task 5: Investigating toxic span detection using RoBERTa pre-training. In *SemEval*.

Kline Finley. 2016. Want to save the comments from trolls? do it yourself. Accessed: 2021-04-15.

Sreyan Ghosh and Sonal Kumar. 2021. Cisco at SemEval-2021 Task 5: What's toxic?: Leveraging transformers for multiple toxic span extraction from online comments. In *SemEval*.

Phu Gia Hoang and Luan Thanh Nguyen. 2021. UIT-E10dot3 at SemEval 2021 Task 5: Toxic spans detection with roberta and spacy's library base systems. In *SemEval*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *ACL*, pages 537–546.

Birol Kuyumcu, Selman Delil, and Cüneyt aksakallı. 2021. Sefamerve_arge at SemEval-2021 Task 5: Toxic span detection using segmentation based 1-d convolutional neural network model. In *SemEval*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Son T. Luu and Ngan Nguyen. 2021. UIT-ISE-NLP at SemEval-2021 Task 5: Toxic span detection with BiLSTM - CRF and toxic BERT comment classification. In *SemEval*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *EMNLP-IJCNLP*, pages 5640–5650.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. In *ACL*, pages 4206–4216.

Viet Anh Nguyen, Tam Nguyen, Huy Dao Quang, and Quang Huu Pham. 2021. S-NLP at semeval-2021 task 5: Toxic spans detection. In *SemEval*.

Marco Palomino, Dawid Grad, and James Bedwell. 2021. An ensemble approach to identify toxicity in text. In *SemEval*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *EMNLP*, pages 1125–1135, Copenghagen, Denmark.

Kamil Pluciński and Hanna Klimczak. 2021. GHOST at SemEval-2021 Task 5: Is explanation all you need? In *SemEval*.

Tharindu Ranasinghe, Diptanu Sarkar, Marcos Zampieri, and Alexander Ororbia. 2021. WLV-RIT at SemEval-2021 Task 5: A neural transformer framework for detecting toxic spans. In *SemEval*.

Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the predictions of any classifier. In *SIGKDD*, pages 1135–1144, San Francisco, USA.

Jonathan Rusert. 2021. NLP_UIOWA at Semeval-2021 Task 5: Transferring toxic sets to tag toxic spans. In *SemEval*.

Rafel Palliser Sans and Albert Rial Farràs. 2021. HLE-UPC at SemEval-2021 Task 5: Multi-Depth DistilBERT for toxic spans detection. In *SemEval*.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Tom De Smedt, Pierre Voué, Sylvia Jaki, Melina Röttcher, and Guy De Pauw. 2020. Profanity & offensive words (POW). *Textgain*.

Thakur Ashutosh Suman and Abhinav Jain. 2021. ASTarTwice at SemEval-2021 Task 5: Toxic span detection using RoBERTa-CRF, domain specific pretraining and self-training. In *SemEval*.

Charles Sutton and Andrew McCallum. 2006. An Introduction to Conditional Random Fields for relational learning. *Introduction to statistical relational learning*, 2:93–128.

Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, volume 30.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *WWW*, pages 1391–1399, Perth, Australia.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *SemEval*.

Hongjie Fan Zhen Wang and Junfei Liu. 2021. MedAI at SemEval-2021 Task 5: Start-to-end tagging framework for toxic spans detection. In *SemEval*.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, and Ruifeng Xu. 2021. HITSZ-HLT at SemEval-2021 Task 5: Span-based ensemble model with toxic lexicon. In *SemEval*.