SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)

Marcos Zampieri[1], Preslav Nakov[2], Sara Rosenthal[3], Pepa Atanasova[4], Georgi Karadzhov[5]
Hamdy Mubarak[2], Leon Derczynski[6], Zeses Pitenis[7], Çağrı Çöltekin[8]

[1]Rochester Institute of Technology, USA, [2]Qatar Computing Research Institute, Qatar
[3]IBM Research, USA, [4]University of Copenhagen, Denmark, [5]University of Cambridge, UK
[6]IT University Copenhagen, Denmark, [7]University of Wolverhampton, UK
[8]University of Tübingen, Germany
marcos.zampieri@rit.edu

Abstract

We present the results and the main findings of SemEval-2020 Task 12 on Multilingual Offensive Language Identification in Social Media (OffensEval-2020). The task included three subtasks corresponding to the hierarchical taxonomy of the OLID schema from OffensEval-2019, and it was offered in five languages: Arabic, Danish, English, Greek, and Turkish. OffensEval-2020 was one of the most popular tasks at SemEval-2020, attracting a large number of participants across all subtasks and languages: a total of 528 teams signed up to participate in the task, 145 teams submitted official runs on the test data, and 70 teams submitted system description papers.

1 Introduction

Offensive language is ubiquitous in social media platforms such as Facebook, Twitter, and Reddit, and it comes in many forms. Given the multitude of terms and definitions related to offensive language used in the literature, several recent studies have investigated the common aspects of different abusive language detection tasks (Waseem et al., 2017; Wiegand et al., 2018). One such example is *SemEval-2019 Task 6: OffensEval*[1](Zampieri et al., 2019b), which is the precursor to the present shared task. OffensEval-2019 used the Offensive Language Identification Dataset (OLID), which contains over 14,000 English tweets annotated using a hierarchical three-level annotation schema that takes both the target and the type of offensive content into account (Zampieri et al., 2019a). The assumption behind this annotation schema is that the target of offensive messages is an important variable that allows us to discriminate between, e.g., hate speech, which often consists of insults targeted toward a *group*, and cyberbullying, which typically targets *individuals*. A number of recently organized related shared tasks followed similar hierarchical models. Examples include HASOC-2019 (Mandl et al., 2019) for English, German, and Hindi, HatEval-2019 (Basile et al., 2019) for English and Spanish, GermEval-2019 for German (Struß et al., 2019), and TRAC-2020 (Kumar et al., 2018b) for English, Bengali, and Hindi.

OffensEval-2019 attracted nearly 800 team registrations and received 115 official submissions, which demonstrates the interest of the research community in this topic. Therefore, we organized a follow-up, OffensEval-2020[2] (SemEval-2020 Task 12), which is described in this report, building on the success of OffensEval-2019 with several improvements. In particular, we used the same three-level taxonomy to annotate new datasets in five languages, where each level in this taxonomy corresponds to a subtask in the competition:

- Subtask A: Offensive language identification;
- Subtask B: Automatic categorization of offense types;
- Subtask C: Offense target identification.

[1]http://sites.google.com/site/offensevalsharedtask/offenseval2019
[2]http://sites.google.com/site/offensevalsharedtask/home

The contributions of OffensEval-2020 can be summarized as follows:

- We provided the participants with a new, large-scale semi-supervised training dataset containing over nine million English tweets (Rosenthal et al., 2020).
- We introduced multilingual datasets, and we expanded the task to four new languages: Arabic (Mubarak et al., 2020b), Danish (Sigurbergsson and Derczynski, 2020), Greek (Pitenis et al., 2020), and Turkish (Çöltekin, 2020). This opens the possibility for cross-lingual training and analysis, which several participants indeed explored.
- Compared to OffensEval-2019, we used larger test datasets for all subtasks.

Overall, OffensEval-2020 was a very successful task. The huge interest demonstrated last year continued this year, with 528 teams signing up to participate in the task, and 145 of them submitting official runs on the test dataset. Furthermore, OffensEval-2020 received 70 system description papers, which is an all-time record for a SemEval task.

The remainder of this paper is organized as follows: Section 2 describes the annotation schema. Section 3 presents the five datasets that we used in the competition. Sections 4-9 present the results and discuss the approaches taken by the participating systems for each of the five languages. Finally, Section 10 concludes and suggests some possible directions for future work.

2 Annotation Schema

OLID's annotation schema proposes a hierarchical modeling of offensive language. It classifies each example using the following three-level hierarchy:

Level A - Offensive Language Detection
Is the text offensive (OFF) or not offesive (NOT)?

NOT: text that is neither offensive, nor profane;

OFF: text containing inappropriate language, insults, or threats.

Level B - Categorization of Offensive Language
Is the offensive text targeted (TIN) or untargeted (UNT)?

TIN: targeted insults or threats towards a group or an individual;

UNT: untargeted profanity or swearing.

Level C - Offensive Language Target Identification
Who or what is the target of the offensive content?

IND: the target is an individual, which can be explicitly mentioned or it can be implicit;

GRP: the target is a group of people based on ethnicity, gender, sexual orientation, religious belief, or other common characteristic;

OTH: the target does not fall into any of the previous categories, e.g., organizations, events, and issues.

3 Data

In this section, we describe the datasets for all five languages: Arabic, Danish, English, Greek, and Turkish. All of the languages follow the OLID annotation schema and all datasets were pre-processed in the same way, e.g., all user mentions were substituted by `@USER` for anonymization. The introduction of new languages using a standardized schema with the purpose of detecting offensive and targeted speech should improve dataset consistency. This strategy is in line with current best practices in abusive language data collection (Vidgen and Derczynski, 2020). All languages contain data for subtask A, and only English contains data for subtasks B and C. The distribution of the data across categories for all languages for subtask A is shown in Table 1, while Tables 2 and 3 present statistics about the data for the English subtasks B and C, respectively. Labeled examples from the different datasets are shown in Table 4.

	Training			Test		
Language	OFF	NOT	Total	OFF	NOT	Total
English	1 448 861	7 640 279	9 089 140	1 090	2 807	3 897
Arabic	1 589	6 411	8 000	402	1 598	2 000
Danish	384	2 577	2 961	41	288	329
Greek	2 486	6 257	8 743	425	1 119	1 544
Turkish	6 131	25 625	31 756	716	2 812	3 528

Table 1: Subtask A (all languages): statistics about the data.

	Training			Test		
Language	TIN	UNT	Total	TIN	UNT	Total
English	149 550	39 424	188 974	850	1 072	1 922

Table 2: Subtask B (English): statistics about the data.

	Training				Test			
Language	IND	GRP	OTH	Total	IND	GRP	OTH	Total
English	120 330	22 176	7 043	149 549	580	190	80	850

Table 3: Subtask C (English): statistics about the data.

Language	**Tweet**	**A**	**B**	**C**
English	This account owner asks for people to think rationally.	NOT	—	—
Arabic	لعنة الله عليك يا سباك يا جبان يابن الكلب.	OFF	—	—
	Translation: May God curse you, O coward, O son of a dog.			
Danish	Du glemmer Østeuropaer som er de værste	OFF	—	—
	Translation: You forget Eastern Europeans, who are the worst			
Greek	Παραδέξου το, είσαι αγάμητη εδώ και καιρό...	OFF	—	—
	Translation: Admit it, you've been unfucked for a while now...			
Turkish	Böyle devam et seni gerizekalı	OFF	—	—
	Translation: Go on like this, you idiot			
English	this job got me all the way fucked up real shit	OFF	UNT	—
English	wtf ari her ass tooo big	OFF	TIN	IND
English	@USER We are a country of morons	OFF	TIN	GRP

Table 4: Annotated examples for all subtasks and languages.

English For English, we provided two datasets: OLID from OffensEval-2019 (Zampieri et al., 2019a), and SOLID, which is a new dataset we created for the task (Rosenthal et al., 2020). SOLID is an abbreviation for Semi-Supervised Offensive Language Identification Dataset, and it contains 9,089,140 English tweets, which makes it the largest dataset of its kind. For SOLID, we collected random tweets using the 20 most common English stopwords such as *the*, *of*, *and*, *to*, etc. Then, we labeled the collected tweets in a semi-supervised manner using democratic co-training, with OLID as a seed dataset. For the co-training, we used four models with different inductive biases: PMI (Turney and Littman, 2003), FastText (Joulin et al., 2017), LSTM (Hochreiter and Schmidhuber, 1997), and BERT (Devlin et al., 2019). We selected the OFF tweets for the test set using this semi-supervised process and we then annotated them manually for all subtasks. We further added 2,500 NOT tweets using this process without further annotation. We computed a Fleiss' κ Inter-Annotator Agreement (IAA) on a small subset of instances that were predicted to be OFF, and obtained 0.988 for Level A (almost perfect agreement), 0.818 for Level B (substantial agreement), and 0.630 for Level C (moderate agreement). The annotation for Level C was more challenging as it is 3-way and also as sometimes there could be different types of targets mentioned in the offensive tweet, but the annotators were forced to choose only one label.

Arabic The Arabic dataset consists of 10,000 tweets collected in April–May 2019 using the Twitter API with the language filter set to Arabic: `lang:ar`. In order to increase the chance of having offensive content, only tweets with two or more vocative particles (*yA* in Arabic) were considered for annotation; the vocative particle is used mainly to direct the speech to a person or to a group, and it is widely observed in offensive communications in almost all Arabic dialects. This yielded 20% offensive tweets in the final dataset. The tweets were manually annotated (for Level A only) by a native speaker familiar with several Arabic dialects. A random subsample of offensive and non-offensive tweets were doubly annotated and the Fleiss κ IAA was found to be 0.92. More details can be found in (Mubarak et al., 2020b).

Danish The Danish dataset consists of 3,600 comments drawn from Facebook, Reddit, and a local newspaper, Ekstra Bladet[3]. The selection of the comments was partially seeded using abusive terms gathered during a crowd-sourced lexicon compilation; in order to ensure sufficient data diversity, this seeding was limited to half the data only. The training data was not divided into distinct training/development splits, and participants were encouraged to perform cross-validation, as we wanted to avoid issues that fixed splits can cause (Gorman and Bedrick, 2019). The annotation (for Level A only) was performed at the individual comment level by males aged 25-40. A full description of the dataset and an accompanying data statement (Bender and Friedman, 2018) can be found in (Sigurbergsson and Derczynski, 2020).

Greek The Offensive Greek Twitter Dataset (OGTD) used in this task is a compilation of 10,287 tweets. These tweets were sampled using popular and trending hashtags, including television programs such as series, reality and entertainment shows, along with some politically related tweets. Another portion of the dataset was fetched using pejorative terms and "you are" as keywords. This particular strategy was adopted with the hypothesis that TV and politics would gather a handful of offensive posts, along with tweets containing vulgar language for further investigation. A team of volunteer annotators participated in the annotation process (for Level A only), with each tweet being judged by three annotators. In cases of disagreement, labels with majority agreement above 66% were selected as the actual tweet labels. The IAA was 0.78 (using Fleiss' κ coefficient). A full description of the dataset collection and annotation is detailed in (Pitenis et al., 2020).

Turkish The Turkish dataset consists of over 35,000 tweets sampled uniformly from the Twitter stream and filtered using a list of the most frequent words in Turkish, as identified by Twitter. The tweets were annotated by volunteers (for Level A only). Most tweets were annotated by a single annotator. The Cohen's κ IAA calculated on 5,000 doubly-annotated tweets was 0.761. Note that we did not include any specific method for spotting offensive language, e.g., filtering by offensive words, or following usual targets of offensive language. As a result, the distribution closely resembles the actual offensive language use on Twitter, with more non-offensive tweets than offensive tweets. More details about the sampling and the annotation process can be found in (Çöltekin, 2020).

4 Task Participation

A total of 528 teams signed up to participate in the task, and 145 of them submitted results: 6 teams made submissions for all five languages, 19 did so for four languages, 11 worked on three languages, 13 on two languages, and 96 focused on just one language. Tables 13, 14, and 15 show a summary of which team participated in which task. A total of 70 teams submitted system description papers, which are listed in Table 12. Below, we analyze the representation and the models used for all language tracks.

Representation The vast majority of teams used some kind of pre-trained embeddings such as contextualized Transformers (Vaswani et al., 2017) and ELMo (Peters et al., 2018) embeddings. The most popular Transformers were BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and the multi-lingual mBERT (Devlin et al., 2019).[4]

[3] `http://ekstrabladet.dk/`

[4] Note that there are some issues with the way mBERT processes some languages, e.g., there is no word segmentation for Arabic, the Danish å/aa mapping is not handled properly (Strømberg-Derczynski et al., 2020), etc.

Many teams also used context-independent embeddings from word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), including language-specific embeddings such as Mazajak (Farha and Magdy, 2019) for Arabic. Some teams used other techniques: word n-grams, character n-grams, lexicons for sentiment analysis, and lexicon of offensive words. Other representations included emoji priors extracted from the weakly supervised SOLID dataset for English, and sentiment analysis using NLTK (Bird et al., 2009), Vader (Hutto and Gilbert, 2014), and FLAIR (Akbik et al., 2018).

Machine learning models In terms of machine learning models, most teams used some kind of pre-trained Transformers: typically BERT, but RoBERTa, XLM-RoBERTa (Conneau et al., 2020), ALBERT (Lan et al., 2019), and GPT-2 (Radford et al., 2019) were also popular. Other popular models included CNNs (Fukushima, 1980), RNNs (Rumelhart et al., 1986), and GRUs (Cho et al., 2014). Older models such as SVMs (Cortes and Vapnik, 1995) were also used, typically as part of ensembles.

5 English Track

A total of 87 teams made submissions for the English track (23 of them participated in the 2019 edition of the task): 27 teams participated in all three English subtasks, 18 teams participated in two English subtasks, and 42 focused on one English subtask only.

Pre-processing and normalization Most teams performed some kind of pre-processing (67 teams) or text normalization (26 teams), which are typical steps when working with tweets. Text normalization included various text transformations such as converting emojis to plain text,[5] segmenting hashtags,[6] general tweet text normalization (Satapathy et al., 2019), abbreviation expansion, bad word replacement, error correction, lowercasing, stemming, and/or lemmatization. Other techniques included the removal of @user mentions, URLs, hashtags, emojis, emails, dates, numbers, punctuation, consecutive character repetitions, offensive words, and/or stop words.

Additional data Most teams found the weakly supervised SOLID dataset useful, and 58 teams ended up using it in their systems. Another six teams gave it a try, but could not benefit from it, and the remaining teams only used the manually annotated training data. Some teams used additional datasets from HASOC-2019 (Mandl et al., 2019), the Kaggle competitions on Detecting Insults in Social Commentary[7] and Toxic Comment Classification[8], the TRAC-2018 shared task on Aggression Identification (Kumar et al., 2018a; Kumar et al., 2018c), the Wikipedia Detox dataset (Wulczyn et al., 2017), and the datasets from (Davidson et al., 2017) and (Wulczyn et al., 2017), as well as some lexicons such as HurtLex (Bassignana et al., 2018) and Hatebase.[9] Finally, one team created their own dataset.

5.1 Subtask A

A total of 82 teams made submissions for subtask A, and the results can be seen in Table 5. This was the most popular subtask among all subtasks and across all languages. The best team UHH-LT achieved an F1 score of 0.9204 using an ensemble of ALBERT models of different sizes. The team ranked second was UHH-LT with an F1 score of 0.9204, and it used RoBERTa-large that was fine-tuned on the SOLID dataset in an unsupervised way, i.e., using the MLM objective. The third team, Galileo, achieved an F1 score of 0.9198, using an ensemble that combined XLM-RoBERTa-base and XLM-RoBERTa-large trained on the subtask A data for all languages. The top-10 teams used BERT, RoBERTa or XLM-RoBERTa, sometimes as part of ensembles that also included CNNs and LSTMs (Hochreiter and Schmidhuber, 1997). Overall, the competition for this subtask was very strong, and the scores are very close: the teams ranked 2–16 are within one point in the third decimal place, and those ranked 2–59 are within two absolute points in the second decimal place from the best team. All but one team beat the majority class baseline (we suspect that team might have accidentally flipped their predicted labels).

[5] http://github.com/carpedm20/emoji
[6] http://github.com/grantjenks/python-wordsegment
[7] http://www.kaggle.com/c/detecting-insults-in-social-commentary
[8] http://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[9] http://hatebase.org/

#	Team	Score	#	Team	Score	#	Team	Score
1	UHH-LT	0.9204	29	UTFPR	0.9094	57	OffensSzeged	0.9032
2	Galileo	0.9198	30	IU-UM@LING	0.9094	58	aprosio	0.9032
3	Rouges	0.9187	31	TAC	0.9093	59	RGCL	0.9006
4	GUIR	0.9166	32	SSN_NLP	0.9092	60	byteam	0.8994
5	KS@LTH	0.9162	33	Hitachi	0.9091	61	ANDES	0.8990
6	kungfupanda	0.9151	34	kathrync	0.9091	62	PUM	0.8973
7	TysonYU	0.9146	35	XD	0.9090	63	shardul007	0.8927
8	AlexU-BackTranslation-TL	0.9139	36	UoB	0.9090	64	I2C	0.8919
9	SpurthiAH	0.9136	37	PAI-NLP	0.9089	65	sonal.kumari	0.8900
10	amsqr	0.9135	38	PingANPAI	0.9089	66	IJS	0.8887
11	m20170548	0.9134	39	VerifiedXiaoPAI	0.9089	67	IR3218-UI	0.8843
12	Coffee_Latte	0.9132	40	nlpUP	0.9089	68	TeamKGP	0.8822
13	wac81	0.9129	41	NLP_Passau	0.9088	69	UNT Linguistics	0.8820
14	hwijeen	0.9129	42	TheNorth	0.9087	70	janecek1	0.8744
15	UJNLP	0.9128	43	problemConquero	0.9085	71	Team Oulu	0.8655
16	ARA	0.9119	44	Lee	0.9084	72	TECHSSN	0.8655
17	Ferryman	0.9115	45	Wu427	0.9081	73	KDELAB	0.8653
18	ALT	0.9114	46	ITNLP	0.9081	74	HateLab	0.8617
19	SINAI	0.9105	47	Better Place	0.9077	75	IASBS	0.8577
20	MindCoders	0.9105	48	IIITG-ADBU	0.9075	76	IUST	0.8288
21	IRLab_DAIICT	0.9104	49	‘doxaAI	0.9075	77	Duluth	0.7714
22	erfan	0.9103	50	NTU_NLP	0.9067	78	RTNLU	0.7665
23	Light	0.9103	51	FERMI	0.9065	79	KarthikaS	0.6351
24	KAFK	0.9099	52	mdherath	0.9063	80	Bodensee	0.4954
25	PALI	0.9098	53	INGEOTEC	0.9061		**Majority Baseline**	0.4193
26	PRHLT-UPV	0.9097	54	PGSG	0.9060	81	IRlab@IITV	0.0728
27	YNU_oxz	0.9097	55	SRIB2020	0.9048			
28	IITP-AINLPML	0.9094	56	GruPaTo	0.9036			

Table 5: Results for English subtask A, ordered by macro-averaged F1 in descending order.

5.2 Subtask B

A total of 41 teams made submissions for subtask B, and the results can be seen in Table 6. The best team is Galileo (which were third on subtask A), whose ensemble model achieved an F1 score of 0.7462. The second-place team, PGSG, used a complex teacher-student architecture built on top of a BERT-LSTM model, which was fine-tuned on the SOLID dataset in an unsupervised way, i.e., optimizing for the MLM objective. NTU_NLP was ranked third with an F1 score of 0.6906. They tackled subtasks A, B, and C as part of a multi-task BERT-based model. Overall, the differences in the scores for subtask B are much larger than for subtask A. For example, the 4th team is two points behind the third one and seven points behind the first one. The top-ranking teams used BERT-based Transformer models, and all but four teams could improve over the majority class baseline.

5.3 Subtask C

A total of 37 teams made submissions for subtask C and the results are shown in Table 7. The best team was once again Galileo, with an F1 score of 0.7145. LT@Helsinki was ranked second with an F1 score of 0.6700. They used fine-tuned BERT with oversampling to improve class imbalance. The third best system was PRHLT-UPV with an F1 score of 0.6692, which combines BERT with hand-crafted features; it is followed very closely by UHH-LT at rank 4, which achieved an F1 score of 0.6683. This subtask is also dominated by BERT-based models, and all teams outperformed the majority class baseline.

Note that the absolute F1-scores obtained by the best teams in the English subtasks A and C are substantially higher than the scores obtained by the best teams in OffensEval-2019: 0.9223 vs. 0.8290 for subtask A and 0.7145 vs. 0.6600 for subtask C. This suggests that the much larger SOLID dataset made available in OffensEval-2020 helped the models make more accurate predictions.

#	Team	Score	#	Team	Score	#	Team	Score
1	Galileo	0.7462	15	Wu427	0.6208	29	PALI	0.5533
2	PGSG	0.7362	16	UNT Linguistics	0.6174	30	HoangDung	0.5524
3	NTU_NLP	0.6906	17	I2C	0.6012	31	KAFK	0.5518
4	UoB	0.6734	18	PRHLT-UPV	0.5987	32	PAI-NLP	0.5451
5	TysonYU	0.6687	19	SRIB2020	0.5805	33	VerifiedXiaoPAI	0.5451
6	GUIR	0.6650	20	FERMI	0.5804	34	Duluth	0.5382
7	UHH-LT	0.6598	21	IU-UM@LING	0.5746	35	Bodensee	0.4926
8	Ferryman	0.6576	22	PingANPAI	0.5687	36	TECHSSN	0.3894
9	IIITG-ADBU	0.6528	23	nlpUP	0.5687	37	KarthikaS	0.3741
10	kathrync	0.6445	24	Team Oulu	0.5676		**Majority Baseline**	0.3741
11	IRLab_DAIICT	0.6412	25	KDELAB	0.5638	38	IRlab@IITV	0.2950
12	INGEOTEC	0.6321	26	wac81	0.5627	39	SSN_NLP	0.2912
13	HateLab	0.6303	27	IITP-AINLPML	0.5569	40	IJS	0.2841
14	AlexU-BackTranslation-TL	0.6300	28	problemConquero	0.5569	41	KEIS@JUST	0.2777

Table 6: Results for English subtask B, ordered by macro-averaged F1 in descending order.

#	Team	Score	#	Team	Score	#	Team	Score
1	Galileo	0.7145	14	KAFK	0.6168	27	nlpUP	0.5515
2	LT@Helsinki	0.6700	15	ssn_nlp	0.6116	28	IS	0.5355
3	PRHLT-UPV	0.6692	16	IJS	0.6094	29	sonal.kumari	0.5260
4	UHH-LT	0.6683	17	PALI	0.6015	30	SRIB2020	0.5147
5	ITNLP	0.6543	18	FERMI	0.5882	31	KEIS@JUST	0.4817
6	wac81	0.6489	19	problemConquero	0.5871	32	ultraviolet	0.4776
7	PUM	0.6473	20	Ferryman	0.5809	33	HateLab	0.4535
8	PingANPAI	0.6394	21	AlexU-BackTranslation-TL	0.5761	34	Bodensee	0.3462
9	IITP-AINLPML	0.6388	22	IIITG-ADBU	0.5756	35	Team Oulu	0.3220
10	PAI-NLP	0.6347	23	Duluth	0.5744	36	SSN_NLP	0.3178
11	GUIR	0.6319	24	KDELAB	0.5720		**Majority Baseline**	0.2704
12	IU-UM@LING	0.6265	25	NTU_NLP	0.5695			
13	mdherath	0.6232	26	INGEOTEC	0.5626			

Table 7: Results for English subtask C, ordered by macro-averaged F1 in descending order.

Furthermore, it suggests that the weakly supervised method used to compile and annotate SOLID is a viable alternative to popular purely manual annotation approaches. A more detailed analysis of the systems' performances will be carried out in order to determine the contribution of the SOLID dataset for the results.

5.4 Best Systems

We provide some more details about the approaches used by the top teams for each subtask. We use subindices to show their rank for each subtask. Additional summaries for some of the best teams can be found in Appendix A.

Galileo (A:3,B:1,C:1) This team was ranked 3rd, 1st, and 1st on the English subtasks A, B, and C, respectively. This is also the only team ranked among the top-3 across all languages. For subtask A, they used multi-lingual pre-trained Transformers based on XLM-RoBERTa, followed by multi-lingual fine-tuning using the OffensEval data. Ultimately, they submitted an ensemble that combined XLM-RoBERTa-base and XLM-RoBERTa-large, achieving an F1 score of 0.9198. For subtasks B and C, they used knowledge distillation in a teacher-student framework, using Transformers such as ALBERT and ERNIE 2.0 (Sun et al., 2020) as teacher models, achieving an F1 score of 0.7462 and 0.7145, for subtasks B and C respectively.

#	Team	Score	#	Team	Score	#	Team	Score
1	ALAMIHamza	0.9017	21	SaiSakethAluru	0.8455	41	tharindu	0.7881
2	ALT	0.9016	22	will_go	0.8440	42	PRHLT-UPV	0.7868
3	Galileo	0.8989	23	erfan	0.8418	43	anitasaroj	0.7793
4	alisafaya	0.8972	24	ANDES	0.8402	44	yemen2016	0.7721
5	AMR-KELEG	0.8958	25	Bushr	0.8395	45	saroarj	0.7474
6	KS@LTH	0.8902	26	klaralang	0.8241	46	kxkajava	0.7306
7	iaf7	0.8778	27	zoher_orabe	0.8221	47	frankakorpel	0.7251
8	sabino	0.8744	28	mircea.tanase	0.8220	48	COMA	0.5436
9	aialharbi	0.8714	29	machouz	0.8216	49	JCT	0.4959
10	yasserotiefy	0.8691	30	orabia	0.8198	50	aprosio	0.4642
11	SAJA	0.8655	31	Taha	0.8183	51	sonal.kumari	0.4536
12	Ferryman	0.8592	32	hamadanayel	0.8182	52	sayanta95	0.4466
13	SAFA	0.8555	33	kathrync	0.8176	53	SpurthiAH	0.4451
14	hhaddad	0.8520	34	fatemah	0.8147		**Majority Baseline**	0.4441
15	TAC	0.8519	35	jbern	0.8125			
16	saradhix	0.8500	36	zahra.raj	0.8057			
17	lukez	0.8498	37	I2C	0.8056			
18	tanvidadu	0.8480	38	jlee24282	0.8024			
19	TysonYU	0.8474	39	problemConquero	0.8021			
20	hwijeen	0.8455	40	asking28	0.8002			

Table 8: Results for Arabic subtask A, ordered by macro-averaged F1 in descending order.

UHH-LT (A:1) This team was ranked 1st on subtask A with an F1 score of 0.9223. They fine-tuned different Transformer models on the OLID training data, and then combined them into an ensemble. They experimented with BERT-base and BERT-large (uncased), RoBERTa-base and RoBERTa-large, XLM-RoBERTa, and four different ALBERT models (large-v1, large-v2, xxlarge-v1, and xxlarge-v2). In their official submission, they used an ensemble combining different ALBERT models. They did not use the labels of the SOLID dataset, but found the tweets it contained nevertheless useful for unsupervised fine-tuning (i.e., using the MLM objective) of the pre-trained Transformers.

6 Arabic Track

A total of 108 teams registered to participate in the Arabic track, and ultimately 53 teams entered the competition with at least one valid submission. Among them, ten teams participated in the Arabic track only, while the rest participated in other languages in addition to Arabic. This was the second shared task for Arabic after the one at the 4th workshop on Open-Source Arabic Corpora and Processing Tools (Mubarak et al., 2020a), which had different settings and less participating teams.

Pre-processing and normalization Most teams performed some kind of pre-processing or text normalization, e.g., Hamza shapes, Alif Maqsoura, Taa Marbouta, diacritics, non-Arabic characters, etc., and only one team replaced emojis with their textual counter-parts.

6.1 Results

Table 8 shows the teams and the F1 scores they achieved for the Arabic subtask A. The majority class baseline had an F1 score of 0.4441, and several teams achieved results that doubled that baseline score. The best-performing team was ALAMIHamza with an F1 score of 0.9017. The second-best team, ALT, was almost tied with the winner, with an F1 score of 0.9016. The Galileo team was third with an F1 score of 0.8989. A summary of the approaches taken by the top-performing teams can be found in Appendix A; here we briefly describe the winning system:

ALAMIHamza (A:1) The winning team achieved the highest F1-score using BERT to encode Arabic tweets, followed by a sigmoid classifier. They further performed translation of the meaning of emojis.

#	Team	Score	#	Team	Score	#	Team	Score
1	LT@Helsinki	0.8119	14	Rouges	0.7587	27	TeamKGP	0.6973
2	Galileo	0.8021	14	Smatgrisene	0.7587	28	Stormbreaker	0.6842
3	NLPDove	0.7923	16	machouz	0.7561	29	TAC	0.6819
4	aprosio	0.7766	17	IU-UM@LING	0.7553	30	Sonal	0.6711
5	KS@LTH	0.7750	18	Ferryman	0.7525	31	RGCL	0.6556
6	JCT	0.7741	19	MindCoders	0.7380	32	PRHLT-UPV	0.6369
7	ANDES	0.7723	20	ARA	0.7267	33	IUST	0.6226
8	TysonYU	0.7685	21	INGEOTEC	0.7237	34	SRIB2020	0.6127
8	FERMI	0.7685	22	KUISAIL	0.7231	35	IR3218-UI	0.5736
10	NLP_Passau	0.7673	23	JAK	0.7086	36	SSN_NLP	0.5678
11	GruPaTo	0.7620	24	LIIR	0.7019	37	Team Oulu	0.5587
12	KEIS@JUST	0.7612	25	MeisterMorxrc	0.6998	38	IJS	0.4913
13	will_go	0.7596	26	problemConquero	0.6974		**Majority Baseline**	0.4668

Table 9: Results for Danish subtask A, ordered by macro-averaged F1 in descending order.

7 Danish Track

A total of 72 teams registered to participate in the Danish track, and 39 of them actually made official submissions on the test dataset. This is the first shared task on offensive language identification to include Danish, and the dataset provided to the OffensEval-2020 participants is an extended version of the one from (Sigurbergsson and Derczynski, 2020).

Pre-processing and normalization Many teams used the pre-processing included in the relevant embedding model, e.g., BPE (Heinzerling and Strube, 2018) and WordPiece. Other pre-processing techniques included emoji normalization, spelling correction, sentiment tagging, lexical and regex-based term and phrase flagging, and hashtag segmentation.

7.1 Results

The results are shown in Table 9. We can see that all teams managed to outperform the majority class baseline. Moreover, all but one team improved over a FastText baseline (F1 = 0.5148), and most teams achieved an F1 score of 0.7 or higher. Interestingly, one of the top-ranked teams, JCT, was entirely non-neural.

LT@Helsinki (A:1) The winning team LT@Helsinki used NordicBERT for representation, as provided by BotXO.[10] NordicBERT is customized to Danish, and avoids some of the pre-processing noise and ambiguity introduced by other popular BERT implementations. The team further reduced orthographic lengthening to maximum two repeated characters, converted emojis to sentiment scores, and used co-occurrences of hashtags and references to usernames. They tuned the hyper-parameters of their model using 10-fold cross validation.

8 Greek Track

A total of 71 teams registered to participate in the Greek track, and ultimately 37 of them made an official submission on the test dataset. This is the first shared task on offensive language identification to include Greek, and the dataset provided to the OffensEval-2020 participants is an extended version of the one from (Pitenis et al., 2020).

Pre-processing and normalization The participants experimented with various pre-processing and text normalization techniques, similarly to what was done for the other languages above. One team further reported replacement of emojis with their textual equivalent.

[10] See http://github.com/botxo/nordic_bert

#	Team	Score	#	Team	Score	#	Team	Score
1	NLPDove	0.8522	14	kathrync	0.8147	27	IUST	0.7756
2	Galileo	0.8507	15	TAC	0.8141	28	KEIS@JUST	0.7730
3	KS@LTH	0.8481	16	IU-UM@LING	0.8140	29	aprosio	0.7700
4	KUISAIL	0.8432	17	MindCoders	0.8137	30	Team Oulu	0.7615
5	IJS	0.8329	18	RGCL	0.8135	31	JCT	0.7568
6	SU-NLP	0.8317	19	problemConquero	0.8115	32	IRlab@IITV	0.7181
7	LT@Helsinki	0.8258	20	Rouges	0.8030	33	TeamKGP	0.7041
8	FERMI	0.8231	21	TysonYU	0.8022	34	SSN_NLP	0.6779
9	Ferryman	0.8222	22	Sonal	0.8017	35	fatemah	0.6036
10	INGEOTEC	0.8197	23	JAK	0.7956	36	CyberTronics	0.4265
11	will_go	0.8176	24	ARA	0.7828		**Majority Baseline**	0.4202
12	ANDES	0.8153	25	machouz	0.7820	37	Stormbreaker	0.2688
13	LIIR	0.8148	26	PRHLT-UPV	0.7763			

Table 10: Results for Greek subtask A, ordered by macro-averaged F1 in descending order.

8.1 Results

The evaluation results are shown in Table 10. The top team, NLPDove, achieved an F1 score of 0.852, with Galileo coming close at the second place with an F1 score of 0.851. The KS@LTH team was ranked third with an F1 score of 0.848. It is no surprise that the majority of the high-ranking submissions and participants used large-scale pre-trained Transformers, with BERT being the most prominent among them, along with wordwvec-style non-contextualized pre-trained word embeddings.

NLPDove (A:1) The winning team NLPDove used pre-trained word embeddings from mBERT, which they fine-tuned using the training data. A domain-specific vocabulary was generated by running the WordPiece algorithm (Schuster and Nakajima, 2012) and using embeddings for extended vocabulary to pre-train and fine-tune the model.

9 Turkish Track

A total of 86 teams registered to participate in the Turkish track, and ultimately 46 of them made an official submission on the test dataset. All teams except for one participated in at least one other track. This is the first shared task on offensive language identification to include Turkish, and the dataset provided to the OffensEval-2020 participants is an extended version of the one from (Çöltekin, 2020).

9.1 Results

The results are shown in Table 11. We can see that team Galileo achieved the highest macro-averaged F1 score of 0.8258, followed by SU-NLP and KUI-SAIL with F1 scores of 0.8167 and 0.8141, respectively. Note that the latter two teams are from Turkey, and they used some language-specific resources and tuning. Most results were in the interval 0.7–0.8, and almost all teams managed to outperform the majority class baseline, which had an F1 score of 0.4435.

Galileo (A:1) The best team in the Turkish subtask A was Galileo, which achieved top results in several other tracks. Unlike the systems ranked second and third, Galileo's system is language-agnostic, and it used data for all five languages in a multi-lingual training setup.

10 Conclusion and Future Work

We presented the results of OffensEval-2020, which featured datasets in five languages: Arabic, Danish, English, Greek, and Turkish. For English, we had three subtasks, representing the three levels of the OLID hierarchy. For the other four languages, we had a subtask for the top-level of the OLID hierarchy only. A total of 528 teams signed up to participate in OffensEval-2020, and 145 of them actually submitted results across all languages and subtasks.

#	Team	Score	#	Team	Score	#	Team	Score
1	Galileo	0.8258	18	LT@Helsinki	0.7719	35	PRHLT-UPV	0.7127
2	SU-NLP	0.8167	19	NLP_Passau	0.7676	36	SRIB2020	0.6993
3	KUISAIL	0.8141	20	will_go	0.7653	37	Team Oulu	0.6868
4	KS@LTH	0.8101	21	FERMI	0.7578	38	ARA	0.6381
5	NLPDove	0.7967	22	problemConquero	0.7553	39	aprosio	0.6268
6	TysonYU	0.7933	23	pin_cod_	0.7496	40	f_shahaby	0.5730
7	RGCL	0.7859	24	TAC	0.7477	41	CyberTronics	0.5420
8	Rouges	0.7815	25	IUST	0.7476	42	IASBS	0.5362
9	GruPaTo	0.7790	26	alaeddin	0.7473	43	JCT	0.5099
10	MindCoders	0.7789	27	fatemah	0.7469	44	machouz	0.4518
11	INGEOTEC	0.7758	28	kathrync	0.7461	45	jooyeon Lee	0.4435
12	Ferryman	0.7737	29	Sonal	0.7422		**Majority Baseline**	0.4435
13	ANDES	0.7737	30	MeisterMorxrc	0.7398	46	Stormbreaker	0.3109
14	I2C	0.7735	31	JAK	0.7334			
15	IU-UM@LING	0.7729	32	KEIS@JUST	0.7330			
16	IJS	0.7724	33	TeamKGP	0.7301			
17	LIIR	0.7720	34	TOBB ETU	0.7154			

Table 11: Results for Turkish subtask A, ordered by macro-averaged F1 in descending order.

Out of the 145 participating teams, 96 teams participated in one language only, 13 teams participated in two languages, 11 in three languages, 19 in four languages, and 6 teams submitted systems for all five languages. The official submissions per language ranged from 37 (for Greek) to 81 (for English). Finally, 70 of the 145 participating teams submitted system description papers, which is an all-time record.

The wide participation in the task allowed us to compare a number of approaches across different languages and datasets. Similarly to OffensEval-2019, we observed that the best systems for all languages and subtasks used large-scale BERT-style pre-trained Transformers such as BERT, RoBERTa, and mBERT. Unlike 2019, however, the multi-lingual nature of this year's data enabled cross-language approaches, which proved quite effective and were used by some of the top-ranked systems.

In future work, we plan to extend the task in several ways. First, we want to offer subtasks B and C for all five languages from OffensEval-2020. We further plan to add some additional languages, especially under-represented ones. Other interesting aspects to explore are code-mixing, e.g., mixing Arabic script and Latin alphabet in the same Arabic message, and code-switching, e.g., mixing Arabic and English words and phrases in the same message. Last but not least, we plan to cover a wider variety of social media platforms.

Acknowledgements

This research was partly supported by the IT University of Copenhagen's Abusive Language Detection project. It is also supported by the Tanbih project at the Qatar Computing Research Institute, HBKU, which aims to limit the effect of "fake news," propaganda and media bias by making users aware of what they are reading.

References

Hwijeen Ahn, Jimin Sun, Chan Young Park, and Jungyun Seo. 2020. NLPDove at SemEval-2020 Task 12: Improving offensive language detection with cross-lingual transfer. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Hamza Alami, Said Ouatik El Alaoui, Abdessamad Benlahbib, and Noureddine En-nahnahi. 2020. LISAC FSDM-USMBA Team at SemEval 2020 Task 12: Overcoming AraBERT's pretrain-finetune discrepancy for Arabic offensive language identification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Abdullah I. Alharbi and Mark Lee. 2020. BhamNLP at SemEval-2020 Task 12: An ensemble of different word embeddings and emotion transfer learning for Arabic offensive language identification in social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Kovacs Alonso, Saini. 2020. TheNorth at SemEval-2020 Task 12: Hate speech detection using RoBERTa. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Talha Anwar and Omer Baig. 2020. TAC at SemEval-2020 Task 12: Ensembling approach for multilingual offensive language identification in social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Aymé Arango, Juan Manuel Pérez, and Franco Luque. 2020. ANDES at SemEval-2020 Task 12: A single BERT multilingual model for offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Pınar Arslan. 2020. pin_cod_ at SemEval-2020 Task 12: Injecting lexicons into bidirectional long short-term memory networks to detect Turkish offensive tweets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. IIITG-ADBU at SemEval-2020 Task 12: Comparison of BERT and BiLSTM in detecting offensive language. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it)*.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly.

Marcos Boriola and Gustavo Paetzold. 2020. UTFPR at SemEval-2020 Task 12: Identifying offensive tweets with lightweight ensembles. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Camilla Casula, Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2020. DH-FBK at SemEval-2020 Task 12: Using multi-channel BERT for multilingual offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*.

Kathryn Chapman, Johannes Bernhard, and Dietrich Klakow. 2020. CoLi @ UdS at SemEval-2020 Task 12: Offensive tweet detection with ensembling. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Po-Chun Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. NTU_NLP at SemEval-2020 Task 12: Identifying offensive tweets using hierarchical multi-task learning approach. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Davide Colla, Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. GruPaTo at SemEval-2020 Task 12: Retraining mBERT on social media and fine-tuned offensive language models. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Tanvi Dadu and Kartikey Pant. 2020. Team Rouges at SemEval-2020 Task 12: Cross-lingual inductive transfer to detect offensive language. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. 2020. Kungfupanda at SemEval-2020 Task 12: BERT-based multi-task learning for offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Kaushik Amar Das, Arup Baruah, Ferdous Ahmed Barbhuiya, and Kuntal Dey. 2020. KAFK at SemEval-2020 Task 12: Checkpoint ensemble of transformers for hate speech classification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17.

Gretel Liz De la Peña Sarracén and Paolo Rosso. 2020. PRHLT-UPV at SemEval-2020 Task 12: BERT for multilingual offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Xiangjue Dong and Jinho D. Choi. 2020. XD at SemEval-2020 Task 12: Offensive language identification in socialmedia with transformer-based ensemble model. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP)*.

Jared Fromknecht and Alexis Palmer. 2020. UNT Linguistics at OffensEval 2020: Linear SVC with pre-trained word embeddings as document vectors and targeted linguistic features. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Kunihiko Fukushima. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202.

Avishek Garain. 2020. Garain at SemEval-2020 Task 12: Sequence based deep learning for categorizing offensive language in social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Erfan Ghadery and Marie-Francine Moens. 2020. LIIR at SemEval-2020 Task 12: A cross-lingual augmentation approach for multilingual offensive language identification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. IITP-AINLPML at SemEval-2020 Task 12: Offensive tweet identification and target categorization in a multitask environment. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ehab Hamdy, Jelena Mitrović, and Michael Granitzer. 2020. nlpUP at SemEval-2020 Task 12: A blazing fast system for offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Keisuke Hanahata and Masaki Aono. 2020. KDELAB at SemEval-2020 Task 12: A system for estimating aggression of tweets using two layers of BERT features. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Sabit Hassan, Younes Samih, Hamdy Mubarak, and Ahmed Abdelali. 2020. ALT at SemEval-2020 Task 12: Arabic and English offensive language identification in social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Peter Juel Henrichsen and Marianne Rathje. 2020. Smatgrisene at SemEval-2020 Task 12: Offense detection by AI – with a pinch of real I. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Mahen Herath, Thushari Atapattu, Dung Anh Hoang, Christoph Treude, and Katrina Falkner. 2020. Adelaide-CyC at SemEval-2020 Task 12: Ensemble of classifiers for offensive language detection in social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

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

Fatemah Husain, Jooyeon Lee, Samuel Henry, and Ozlem Uzuner. 2020. SalamNET at SemEval-2020 Task 12: Deep learning approach for Arabic offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Omar Hussein, Hachem Sfar, Jelena Mitrović, and Michael Granitzer. 2020. NLP_Passau at SemEval-2020 Task 12: Multilingual neural network for offensive language detection in English, Danish and Turkish. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Clayton J Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2020. AlexU-BackTranslation-TL at SemEval-2020 Task 12: Improving offensive language detection using data augmentation and transfer learning. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Md Saroar Jahan and Mourad Oussalah. 2020. Team Oulu at SemEval-2020 Task 12: Multilingual identification of offensive language, type and target of Twitter post using translated datasets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Piotr Janiszewski, Mateusz Skiba, and Urszula Walińska. 2020. PUM at SemEval-2020 Task 12: aggregation of transformer-based models' features for offensive language recognition. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Li Junyi, Zhou Xiaobing, and Zhang Zichen. 2020. Lee at SemEval-2020 Task 12: A BERT model based on the maximum self-ensemble strategy for identifying offensive. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

A. Kalaivani and D. Thenmozhi. 2020. SSN_NLP_MLRG at SemEval-2020 Task 12: Offensive language identification in English, Danish, Greek using BERT and machine learning approach. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC)*.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018b. Evaluating Aggression Identification in Social Media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, Santa Fe, USA.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018c. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Sonal Kumari. 2020. Sonal.kumari at SemEval-2020 Task 12: Social media multilingual offensive text identification and categorization using neural network models. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Sandy Kurniawan, Indra Budi, and Muhammad Okky Ibrohim. 2020. IR3218-UI at SemEval-2020 Task 12: Emoji effects on offensive language identification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Karishma Laud, Jagriti Singh, Randeep Kumar Sahu, and Ashutosh Modi. 2020. problemConquero at SemEval-2020 Task 12: Transformer and soft label-based approaches. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Wah Meng Lim and Harish Tayyar Madabushi. 2020. UoB at SemEval-2020 Task 6: Boosting BERT with corpus level information. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

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*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC Track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE)*.

Abir Messaoudi, Hatem Haddad, and Moez Ben Haj Hmida. 2020. Compass at SemEval-2020 Task 12: From a syntax-ignorant n-gram embeddings model to a deep bidirectional language model. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Sabino Miranda-Jiménez, Eric S. Tellez, Mario Graff, and Daniela Moctezuma. 2020. INGEOTEC at SemEval-2020 Task 12: Multilingual classification of offensive text. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Alejandro Mosquera. 2020. amsqr at SemEval-2020 Task 12: Offensive language detection using neural networks and anti-adversarial features. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020a. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020b. Arabic offensive language on Twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Hamada A. Nayel. 2020. NAYEL at SemEval-2020 Task 12: TF/IDF-based approach for automatic offensive language detection in Arabic tweets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Zoher Orabe, Bushr Haddad, Nada Ghneim, and Anas Al-Abood. 2020. DoTheMath at SemEval-2020 Task 12: Deep neural networks with self attention for Arabic offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Yasser Otiefy, Ahmed Abdelmalek, and Islam El Hosary. 2020. WOLI at SemEval-2020 Task 12: Arabic offensive language identification on different Twitter datasets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Xiaozhi Ou, Xiaobing Zhou, and Xuejie Zhang. 2020. YNU_oxz at SemEval-2020 Task 12: Bidirectional GRU with capsule for identifying multilingual offensive language. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Apurva Parikh, Abhimanyu Singh Bisht, and Prasenjit Majumder. 2020. IRLab_DAIICT at SemEval-2020 Task 12: Machine learning and deep learning methods for offensive language identification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Ted Pedersen. 2020. Duluth at SemEval-2020 Task 12: Offensive tweet identification in English with logistic regression. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT)*.

Bao-Tran Pham-Hong and Setu Chokshi. 2020. PGSG at SemEval-2020 Task 12: BERT-LSTM with tweets' pretrained model and noisy student training method. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

Flor Miriam Plaza-del Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. 2020. SINAI at SemEval-2020 Task 12: Offensive language identification exploring transfer learning models. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Marc Pàmies, Emily Öhman, Kaisla Kajava, and Jörg Tiedemann. 2020. LT@Helsinki at SemEval-2020 Task 12: Multilingual or language-specific BERT? In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).

Tharindu Ranasinghe and Hansi Hettiarachchi. 2020. BRUMS at SemEval-2020 Task 12: Transformer based multilingual offensive language identification in social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Manikandan Ravikiran, Amin Ekant Muljibhai, Toshinori Miyoshi, Hiroaki Ozaki, Yuta Koreeda, and Sakata Masayuki. 2020. Hitachi at SemEval-2020 Task 12: Offensive language identification with noisy labels using statistical sampling and post-processing. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 Task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Anita Saroj, Supriya Chanda, and Sukomal Pal. 2020. IRlab@IITV at SemEval-2020 Task 12: multilingual offensive language identification in social media using SVM. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Ranjan Satapathy, Yang Li, Sandro Cavallari, and Erik Cambria. 2019. Seq2seq deep learning models for microtext normalization. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Paul Sayanta, Saha Sriparna, and Hasanuzzaman Mohammed. 2020. CyberTronics at SemEval-2020 Task 12: Multilingual offensive language identification over social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

Abhishek Singh and Surya Pratap Singh Parmar. 2020. Voice@SRIB at SemEval-2020 Task [9,12]: Sentiment and offensiveness detection in social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Rajalakshmi Sivanaiah, Angel Deborah S, S Milton Rajendram, and Mirnalinee T T. 2020. TECHSSN at SemEval-2020 Task 12: Offensive language detection using BERT embeddings. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Kasper Socha. 2020. KS@LTH at SemEval-2020 Task 12: Fine-tuning multi- and monolingual transformer models for offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Sajad Sotudeh, Tong Xiang, Hao-Ren Yao, Sean MacAvaney, Eugene Yang, Nazli Goharian, and Ophir Frieder. 2020. GUIR at SemEval-2020 Task 12: Domain-tuned contextualized models for offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Leon Strømberg-Derczynski, Rebekah Baglini, Morten H Christiansen, Manuel R Ciosici, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, et al. 2020. The Danish Gigaword Project. *arXiv preprint arXiv:2005.03521*.

Julia Maria Struß, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of KONVENS*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*.

Shardul Suryawanshi, Mihael Arcan, and Paul Buitelaar. 2020. NUIG at SemEval-2020 Task 12: Deep learning with pseudo labelled data. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Mircea-Adrian Tanase, Dumitru-Clementin Cercel, and Costin-Gabriel Chiru. 2020. UPB at SemEval-2020 Task 12: Multilingual offensive language detection on social media by fine-tuning a variety of BERT-based models. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Saja Khaled Tawalbeh, Mahmoud Hammad, and Mohammad AL-Smadi. 2020. KEIS@JUST at SemEval-2020 Task 12: Identifying multilingual offensive tweets using weighted ensemble and fine-tuned BERT. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Moshe Uzan and HaCohen-Kerner Yaakov. 2020. JCT at SemEval-2020 Task 12: Offensive language detection in tweets using preprocessing methods, character and word n-grams. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems (NIPS)*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670*.

Susan Wang and Zita Marinho. 2020. Nova-Wang at SemEval-2020 Task 12: OffensEmblert: an ensemble of offensive language classifiers. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Shuohuan Wang, Jiaxiang Liu, Xuan Ouyang, and Yu Sun. 2020. Galileo at SemEval-2020 Task 12: Multilingual learning for offensive language identification using pre-trained language models. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*.

Chen Weilong, Wang Peng, Li Jipeng, Zheng Yuanshuai, Wang Yan, and Zhang Yanru. 2020. Ferryman at SemEval-2020 Task 12: BERT-based model with advanced improvement methods for multilingual offensive language identification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Gregor Wiedemann, Seid Yimam, and Chris Biemann. 2020. UHH-LT at SemEval-2020 Task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop (GermEval)*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*.

Yinnan Yao, Nan Su, and Kun Ma. 2020. UJNLP at SemEval-2020 Task 12: Detecting offensive language using bidirectional transformers. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT)*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval)*.

Victoria Pachón Álvarez, Jacinto Mata Vázquez, José Manuel López Betanzos, and José Luis Arjona Fernández. 2020. I2C in SemEval2020 Task 12: Simple but effective approaches to offensive speech detection in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Anıl Özdemir and Reyyan Yeniterzi. 2020. SU-NLP at SemEval-2020 Task 12: Offensive language identification in Turkish tweets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

A Best-Performing Teams

Below we present a short overview of the top-3 systems for all subtasks and for all languages:

Galileo (EN B:1, EN C:1, TR A:1; DK A:2, GR A:2; AR A:3, EN A:3) This team was ranked 3rd, 1st, and 1st on the English subtasks A, B, and C, respectively; it was also ranked 1st for Turkish, 2nd for Greek and 3rd for Arabic and Danish. This is the only team ranked among the top-3 across all languages. For subtask A (all languages), they used multi-lingual pre-trained Transformers based on XLM-RoBERTa, followed by multi-lingual fine-tuning using the OffensEval data. Ultimately, they submitted an ensemble that combined XLM-RoBERTa-base and XLM-RoBERTa-large. For the English subtasks B and C, they used knowledge distillation in a teacher-student framework, using Transformers such as ALBERT and ERNIE 2.0 (Sun et al., 2020) as teacher models.

UHH-LT (EN A:1) This team was ranked 1st on the English subtask A. They fine-tuned different Transformer models on the OLID training data, and then combined them into an ensemble. They experimented with BERT-base and BERT-large (uncased), RoBERTa-base and RoBERTa-large, XLM-RoBERTa, and four different ALBERT models (large-v1, large-v2, xxlarge-v1, and xxlarge-v2). In their official submission, they used an ensemble combining different ALBERT models. They did not use the labels of the SOLID dataset, but found the tweets it contained nevertheless useful for unsupervised fine-tuning (i.e., using the MLM objective) of the pre-trained Transformers.

LT@Helsinki (DK A:1; EN C:2) This team was ranked 1st for Danish and 2nd for English subtask C. For Danish, they used NordicBERT, which is customized to Danish, and avoids some of the pre-processing noise and ambiguity introduced by other popular BERT implementations. The team further reduced orthographic lengthening to maximum two repeated characters, converted emojis to sentiment scores, and used co-occurrences of hashtags and references to usernames. They tuned the hyper-parameters of their model using 10-fold cross validation. For English subtask C, they used a very simple approach: over-sample the training data to overcome the class imbalance, and then fine-tune BERT-base-uncased.

NLPDove (GR A:1; DK A:3) This team was ranked 1st for Greek and 3rd for Danish. This team used extensive preprocessing and two data augmentation strategies: using additional semi-supervised labels from SOLID with different thresholds, and cross-lingual transfer with data selection. They further proposed and used a new metric, Translation Embedding Distance, in order to measure the transferability of instances for cross-lingual data selection. Moreover, they used data from different languages to fine-tune an mBERT model. Ultimately, they used a majority vote ensemble of several mBERT models, with minor variations in the parameters.

ALAMIHamza (AR A:1) This team was ranked 1st for Arabic. They used BERT to encode Arabic tweets, followed by a sigmoid classifier. They further performed translation of the meaning of emojis.

PGSG (EN B:2) The team was ranked 2nd on the English subtask B. They first fine-tuned the BERT-Large, Uncased (Whole Word Masking) checkpoint using the tweets from SOLID, but ignoring their labels. For this, they optimized for the MLM objective only, without the Next Sentence Prediction loss in BERT. Then, they built a BERT-LSTM model using this fine-tuned BERT, and adding LSTM layers on top of it, together with the [CLS] token. Finally, they used this architecture to train a Noisy Student model using the SOLID data.

ALT (AR A:2) The team was ranked 2nd for Arabic. They used an ensemble of SVM, CNN-BiLSTM and Multilingual BERT. The SVMs used character n-grams, word n-grams, word embeddings as features, while the CNN-BiLSTM learned character embeddings and further used pre-trained word embeddings as input.

SU-NLP (TR A:2) The team was ranked 2nd for Turkish. They used an ensemble of three different models: CNN-LSTM, BiLSTM-Attention, and BERT. They further used word embeddings, pre-trained on tweets, and BERTurk, BERT model for Turkish.

Rouges (EN A:3) The team was ranked 3rd for the English subtask A. They used XLM-RoBERTa fine-tuned sequentially on all languages in a particular order: English, then Turkish, then Greek, then Arabic, then Danish.

NTU_NLP (EN B:3) This team was ranked 3rd on the English subtask B. They proposed a hierarchical multi-task learning approach that solves subtasks A, B, and C simultaneously, following the hierarchical structure of the annotation schema of the OLID dataset. Their architecture has three layers. The input of the first layer is the output of BERT, and its output (D1-OUT) is directly connected to the output layer for subtask A. The second layer's input is the BERT output concatenated with D1-OUT, and its output (D2-OUT) is directly connected to the output layer for subtask B. The third layer's input is the BERT output concatenated with D2-OUT, and its output is directly connected to the output layer for subtask C.

PRHLT-UPV (EN C:3) The team was ranked 3rd on the English subtask C. They used a combination of BERT and hand-crafted features, which were concatenated to the [CLS] representation from BERT. The features include the length of the tweets, the number of misspelled words, and the use of punctuation marks, emoticons, and noun phrases.

KS@LTH (GR: A:3) This team was ranked 3rd for Greek. They experimented with monolingual and cross-lingual models, BERT and XLM-RoBERTa model, respectively, and they found BERT to perform slightly better.

KUISAIL (TR: A:3) This team was ranked 3rd for Turkish. They combined BERTurk with a CNN, in a BERT-CNN model.

B Participants

Team	System Description Paper	Team	System Description Paper
AdelaideCyC	(Herath et al., 2020)	LISAC FSDM-USMBA	(Alami et al., 2020)
AlexU-BackTranslation-TL	(Ibrahim et al., 2020)	LT@Helsinki	(Pàmies et al., 2020)
ALT	(Hassan et al., 2020)	NAYEL	(Nayel, 2020)
amsqr	(Mosquera, 2020)	NLP_Passau	(Hussein et al., 2020)
ANDES	(Arango et al., 2020)	NLPDove	(Ahn et al., 2020)
BhamNLP	(Alharbi and Lee, 2020)	nlpUP	(Hamdy et al., 2020)
JCT	(Uzan and Yaakov, 2020)	Nova-Wang	(Wang and Marinho, 2020)
BRUMS	(Ranasinghe and Hettiarachchi, 2020)	NTU_NLP	(Chen et al., 2020)
CoLi @ UdS	(Chapman et al., 2020)	NUIG	(Suryawanshi et al., 2020)
CyberTronics	(Sayanta et al., 2020)	Oulu	(Jahan and Oussalah, 2020)
DoTheMath	(Orabe et al., 2020)	PGSG	(Pham-Hong and Chokshi, 2020)
Duluth	(Pedersen, 2020)	pin_cod_	(Arslan, 2020)
FBK-DH	(Casula et al., 2020)	PRHLT-UPV	(De la Peña Sarracén and Rosso, 2020)
Ferryman	(Weilong et al., 2020)	problemConquero	(Laud et al., 2020)
Galileo	(Wang et al., 2020)	PUM	(Janiszewski et al., 2020)
Garain	(Garain, 2020)	Rouges	(Dadu and Pant, 2020)
GruPaTo	(Colla et al., 2020)	SalamNET	(Husain et al., 2020)
GUIR	(Sotudeh et al., 2020)	SINAI	(Plaza-del Arco et al., 2020)
Hitachi	(Ravikiran et al., 2020)	Smatgrisene	(Henrichsen and Rathje, 2020)
I2C	(Álvarez et al., 2020)	Sonal.kumari	(Kumari, 2020)
iCompass	(Messaoudi et al., 2020)	SRIB2020	(Singh and Parmar, 2020)
IIITG-ADBU	(Baruah et al., 2020)	SSN_NLP_MLRG	(Kalaivani and Thenmozhi, 2020)
IITP-AINLPML	(Ghosh et al., 2020)	SU-NLP	(Özdemir and Yeniterzi, 2020)
INGEOTEC	(Miranda-Jiménez et al., 2020)	TAC	(Anwar and Baig, 2020)
IR3218-UI	(Kurniawan et al., 2020)	TECHSSN	(Sivanaiah et al., 2020)
IRlab@IITV	(Saroj et al., 2020)	TheNorth	(Alonso, 2020)
IRLab_DAIICT	(Parikh et al., 2020)	UHH-LT	(Wiedemann et al., 2020)
KAFK	(Das et al., 2020)	UJNLP	(Yao et al., 2020)
KDELAB	(Hanahata and Aono, 2020)	UNT	(Fromknecht and Palmer, 2020)
KEIS@JUST	(Tawalbeh et al., 2020)	UoB	(Lim and Madabushi, 2020)
KS@LTH	(Socha, 2020)	UPB	(Tanase et al., 2020)
KUISAIL	(Safaya et al., 2020)	UTFPR	(Boriola and Paetzold, 2020)
Kungfupanda	(Dai et al., 2020)	WOLI	(Otiefy et al., 2020)
Lee	(Junyi et al., 2020)	XD	(Dong and Choi, 2020)
LIIR	(Ghadery and Moens, 2020)	YNU_oxz	(Ou et al., 2020)

Table 12: The teams that participated in OffensEval-2020 and submitted system description papers and the corresponding reference to their papers.

Team	A-Arabic	A-Danish	A-Greek	A-Turkish	A-English	B-English	C-English
AlexU-BackTranslation-TL					✔	✔	✔
ALT	✔				✔		
aialharbi	✔						
alaeddin				✔			
ALAMIHamza	✔						
alisafaya	✔						
AMR-KELEG	✔						
amsqr					✔		
ANDES	✔	✔	✔	✔	✔		
anitasaroj	✔						
aprosio	✔	✔	✔	✔	✔		
ARA		✔	✔	✔	✔		
asking28	✔						
Better Place					✔		
Bodensee					✔	✔	✔
Bushr	✔						
byteam					✔		
Coffee_Latte					✔		
COMA	✔						
CyberTronics			✔	✔			
doxaAI					✔		
Duluth					✔	✔	✔
erfan	✔				✔		
f_shahaby				✔			
fatemah	✔		✔	✔			
FERMI		✔	✔	✔	✔	✔	✔
Ferryman	✔	✔	✔	✔	✔	✔	✔
frankakorpel	✔						
Galileo	✔	✔	✔	✔	✔	✔	✔
GruPaTo		✔		✔	✔		
GUIR					✔	✔	✔
hamadanayel	✔						
HateLab					✔	✔	✔
hhaddad	✔						
Hitachi					✔		
HoangDung						✔	
hwijeen	✔				✔		
I2C	✔			✔	✔	✔	
iaf7	✔						
IASBS				✔	✔		
IIITG-ADBU					✔	✔	✔
IITP-AINLPML					✔	✔	✔
IJS		✔	✔	✔	✔	✔	✔
INGEOTEC		✔	✔	✔	✔	✔	✔
IR3218-UI		✔			✔		
IRlab@IITV			✔		✔	✔	
IRLab_DAIICT					✔	✔	
IS							✔
ITNLP					✔		✔
IU-UM@LING		✔	✔	✔	✔	✔	✔
IUST		✔	✔	✔	✔		
JAK		✔	✔	✔			
janecek1					✔		
jbern	✔						
JCT	✔	✔	✔	✔			
jlee24282	✔						
jooyeon Lee				✔			
KAFK					✔	✔	✔
KarthikaS					✔	✔	
kathrync	✔		✔	✔	✔	✔	
KDELAB					✔	✔	✔
KEIS@JUST		✔	✔	✔		✔	✔

Table 13: Overview of team participation in the subtasks (part 1).

Team	A-Arabic	A-Danish	A-Greek	A-Turkish	A-English	B-English	C-English
klaralang	✔						
KS@LTH	✔	✔	✔	✔	✔		
KUISAIL				✔			
kungfupanda					✔		
kxkajava	✔						
Lee					✔		
Light					✔		
LIIR		✔	✔	✔			
LT@Helsinki		✔	✔	✔			✔
lukez	✔						
m20170548					✔		
machouz	✔	✔	✔	✔			
mdherath					✔		✔
MeisterMorxrc		✔		✔			
MindCoders		✔	✔	✔	✔		
mircea.tanase	✔						
NLP_Passau		✔		✔	✔		
NLPDove		✔	✔	✔			
nlpUP					✔	✔	✔
NTU_NLP					✔	✔	✔
OffensSzeged					✔		
orabia	✔						
Oulu				✔			
PAI-NLP					✔	✔	✔
PALI					✔	✔	✔
PGSG					✔	✔	
pin_cod_				✔			
PingANPAI					✔	✔	✔
PRHLT-UPV	✔	✔	✔	✔	✔	✔	✔
problemConquero	✔	✔	✔	✔	✔	✔	✔
PUM					✔		✔
RGCL		✔	✔	✔	✔		
Rouges		✔	✔	✔	✔		
RTNLU					✔		
sabino	✔						
SAFA	✔						
SaiSakethAluru	✔						
SAJA	✔						
saradhix	✔						
saroarj	✔						
sayanta95	✔						
shardul007					✔		
SINAI					✔		
Smatgrisene		✔					
Sonal		✔	✔	✔			
sonal.kumari	✔				✔		✔
SpurthiAH	✔				✔		
SRIB2020		✔		✔	✔	✔	✔
SSN_NLP		✔	✔		✔	✔	✔
Stormbreaker		✔	✔	✔			
SU-NLP			✔	✔			
Taha	✔						
TAC	✔	✔	✔	✔	✔		
tanvidadu	✔						
GruPaTo				✔	✔		
Team Oulu		✔	✔		✔	✔	✔
TeamKGP		✔	✔	✔	✔		
TECHSSN					✔	✔	
tharindu	✔						
TheNorth					✔		
TOBB ETU				✔			
TysonYU	✔	✔	✔	✔	✔	✔	
UHH-LT					✔	✔	✔
UJNLP					✔		

Table 14: Overview of team participation in the subtasks (part 2).

Team	A-Arabic	A-Danish	A-Greek	A-Turkish	A-English	B-English	C-English
ultraviolet							✔
UNT Linguistics					✔	✔	
UoB					✔	✔	
UTFPR					✔		
VerifiedXiaoPAI					✔	✔	
wac81					✔	✔	✔
will_go	✔	✔	✔	✔			
KUISAIL		✔	✔				
Wu427					✔	✔	
XD					✔		
yasserotiefy	✔						
yemen2016	✔						
YNU_oxz					✔		
zahra.raj	✔						
zoher_orabe	✔						

Table 15: Overview of team participation in the subtasks (part 3).