SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)

Marcos Zampieri[1], Shervin Malmasi[2], Preslav Nakov[3], Sara Rosenthal[4]
Noura Farra[5], Ritesh Kumar[6]
[1]University of Wolverhampton, UK, [2]Amazon Research, USA
[3]Qatar Computing Research Institute, HBKU, Qatar
[4]IBM Research, USA, [5]Columbia University, USA, [6]Bhim Rao Ambedkar University, India
m.zampieri@wlv.ac.uk

Abstract

This paper presents the results and main findings of the shared task on Identifying and Categorizing Offensive Language in Social Media (OffensEval). SemEval-2019 Task 6 provided participants with an annotated dataset containing English tweets. The competition was divided into three sub-tasks. In sub-task A systems were trained to discriminate between offensive and non-offensive tweets, in sub-task B systems were trained to identify the type of offensive content in the post, and finally, in sub-task C systems were trained to identify the target of offensive posts. OffensEval attracted a large number of participants and it was one of the most popular tasks in SemEval 2019. In total, nearly 800 teams signed up to participate in the task and 115 of them submitted results which are presented and analyzed in this report.

1 Introduction

The automatic identification of offensive content online is an important task which has gained more attention in recent years. Social media platforms such as Facebook and Twitter have been investing heavily in ways to cope with the widespread forms of such content. The task is usually modelled as supervised classification problem in which systems are trained using a dataset containing posts which are annotated with respect to the presence of some form(s) of abusive or offensive content. Examples of offensive content investigated in previous studies include hate speech (Davidson et al., 2017; Malmasi and Zampieri, 2017, 2018), cyberbulling (Dinakar et al., 2011), and aggression (Kumar et al., 2018).

Given the multitude of terms and definitions used in the literature, recent studies have investigated common aspects of the abusive language detection sub-tasks (Waseem et al., 2017; Wiegand et al., 2018). However, none of these initial studies focused on both the type *and* the target of the offensive language. Therefore, in conjunction with this task, we present the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019). OLID is an annotated dataset with a three-level annotation model. We show that breaking down offensive content into sub-categories by taking the type and target of offenses into account results in a flexible annotation model that can relate to the phenomena captured by previously annotated datasets such as the one by (Davidson et al., 2017). Hate speech, for example, is commonly understood as an insult targeted at a *group* whereas cyberbulling is typically targeted at an *individual*). In OffensEval[1] we use OLID (Zampieri et al., 2019) and propose one sub-task for each layer of annotation as presented in Section 3.

The remainder of this paper is organized as follows: Section 3 presents the shared task description and the sub-tasks included in OffensEval and Section 4 includes a brief description of OLID based on Zampieri et al. (2019). Section 5 presents an analysis of the results of the shared task, and, finally, Section 6 concludes this paper presenting avenues for future work.

2 Related Work

Different abusive and offense language identification sub-tasks have been explored in the past few years including aggression identification, bullying detection, hate speech, toxic comments, and offensive language.

Aggression identification: The TRAC shared task on Aggression Identification (Kumar et al., 2018) provided participants with a dataset containing 15,000 annotated Facebook posts and com-

[1]https://competitions.codalab.org/competitions/20011

ments in English and Hindi for training and validation. For testing, two different sets, one from Facebook and one from Twitter were provided. Systems were trained to discriminate between three classes: non-aggressive, covertly aggressive, and overtly aggressive. The best performing systems in this competition used deep learning approaches based on convolutional neural networks (CNN), recurrent neural networks, and LSTMs (Aroyehun and Gelbukh, 2018; Majumder et al., 2018).

Bullying detection: Several studies have been published on bullying detection. One of them is the one by Xu et al. (2012) which apply sentiment analysis to detect bullying in tweets. Xu et al. (2012) use topic models to to identify relevant topics in bullying. Another related study is the one by Dadvar et al. (2013) which use user-related features such as the frequency of profanity in previous messages to improve bullying detection.

Hate speech identification: It is perhaps the most widespread abusive language detection subtask. There have been several studies published on this sub-task such as Kwok and Wang (2013) and Djuric et al. (2015) who build a binary classifier to distinguish between 'clean' comments and comments containing hate speech and profanity. More recently, Davidson et al. (2017) presented the hate speech detection dataset containing over 24,000 English tweets labeled as non offensive, hate speech, and profanity.

Offensive language: The GermEval[2] (Wiegand et al., 2018) shared task focused on Offensive language identification in German tweets. A dataset of over 8,500 annotated tweets was provided for a course-grained binary classification task in which systems were trained to discriminate between offensive and non-offensive tweets and a second task where the organizers broke down the offensive class into three classes: profanity, insult, and abuse.

Toxic comments: The Toxic Comment Classification Challenge was an open competition at Kaggle which provided participants with comments from Wikipedia labeled in six classes: toxic, severe toxic, obscene, threat, insult, identity hate.

While each of these sub-tasks tackle a particular type of abuse or offense, they share similar properties and the hierarchical annotation model proposed proposed in OLID (Zampieri et al., 2019) and used in OffensEval aims to capture this. Considering that, for example, an insult targeted at an individual is commonly known as cyberbulling and that insults targeted at a group are known as hate speech, we pose that OLID's hierarchical annotation model makes it a useful resource for various offensive language identification sub-tasks.

[2]https://projects.fzai.h-da.de/iggsa/

3 Task Description and Evaluation

The training and testing material used for OffensEval is the aforementioned Offensive Language Identification Dataset (OLID) dataset, built for this task. OLID was annotated using a hierarchical three-level annotation model introduced in Zampieri et al. (2019). We use the annotation of each of the three layers in OLID to each sub-task in OffensEval as follows:

3.1 Sub-task A: Offensive language identification

In this sub-task, systems are trained to discriminate between offensive and non-offensive posts. Offensive posts include insults, threats, and posts containing any form of untargeted profanity. Each instance receives one of the two following labels.

- Not Offensive (NOT): Posts that do not contain offense or profanity;
- Offensive (OFF): We label a post as offensive if it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This category includes insults, threats, and posts containing profane language or swear words.

3.2 Sub-task B: Automatic categorization of offense types

In sub-task B, systems are trained to categorize offenses. Only posts containing the label Offensive (OFF) in sub-task A are included in sub-task B. The two categories in sub-task B are the following:

- Targeted Insult (TIN): Posts containing an insult/threat to an individual, group, or others (see next sub-task);
- Untargeted (UNT): Posts containing non-targeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language.

3.3 Level C: Offense target identification

Sub-task C focuses on the target of offenses. Only posts which are either insults or threats (TIN) received this third layer of annotation. The three labels in sub-task C are the following:

- Individual (IND): Posts targeting an individual. It can be a a famous person, a named individual or an unnamed participant in the conversation. Insults/threats targeted at individuals are often defined as cyberbullying.

- Group (GRP): The target of these offensive posts is a group of people considered as a unity due to the same ethnicity, gender or sexual orientation, political affiliation, religious belief, or other common characteristic. Many of the insults and threats targeted at a group correspond to what is commonly understood as hate speech.

- Other (OTH): The target of these offensive posts does not belong to any of the previous two categories (e.g. an organization, a situation, an event, or an issue).

3.4 Task Evaluation

Given the strong imbalance between the number of instances in each class across the three tasks, we used the macro-averaged F1-score as the official evaluation metric for all tasks. This metric weights precision and recall equally, and calculates the F1-score for each class independently. The values are then averaged, giving equal weight to all classes, regardless of the number of samples.

4 Data

In this Section we summarize OLID, the dataset used for this task. A detailed description of the data collection process and annotation is presented in Zampieri et al. (2019).

OLID is a large collection of English tweets annotated using a hierarchical three-layer annotation model. It contains 14,100 annotated tweets divided in a training partition containing 13,240 tweets and a test partition containing 860 tweets. Additionally, a small trial set containing 320 tweets was made available before the start of the competition.

The distribution of the labels in OLID is shown in Table 1. We annotated the dataset using the crowdsourcing platform Figure Eight.[3] Finally, four examples of annotated instances in the dataset are presented in Table 2.

A	B	C	Train	Test	Total
OFF	TIN	IND	2,407	100	2,507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1,074	78	1,152
OFF	UNT	—	524	27	551
NOT	—	—	8,840	620	9,460
All			13,240	860	14,100

Table 1: Distribution of label combinations in OLID.

5 Results

The models used in the Task submissions ranged from traditional machine learning (e.g. SVM and logistic regression), deep learning (e.g. CNN, RNN, BiLSTM, using attention), to the state-of-the art deep learning language models (e.g. ELMO (Peters et al., 2018), BERT (Devlin et al., 2018)). Figure 1 displays a pie chart indicating the breakdown of model type for all submissions in sub-task A. Deep learning is clearly the most popular approach. Similar trends of model type were seen in sub-tasks B and C.

Several external datasets were explored such as Hate Speech Tweets (Davidson et al., 2017), toxicity labels (Thain et al., 2017), and TRAC (Kumar et al., 2018). Seven systems indicated the use of sentiment lexicons or a sentiment model for prediction and two indicated the use of offensive word lists. In addition, several models explored using pre-trained embeddings such as FastText (Bojanowski et al., 2016), Glove and Twitter Glove (Pennington et al., 2014), Twitter word2vec (Godin et al., 2015).

In addition, many pre-processing techniques were tried such as token, hashtag, url, retweet (RT), and date normalization, elongated words (e.g. "Hiiiii" to "Hi", partially hidden words ("c00l" to "cool") converting emojis to text, removing uncommon words, and Twitter specific tokenizers (Ark Twokenizer[4] (Gimpel et al., 2011), NLTK TweetTokenizer[5]) as well as standard tokenizers (Stanford Core NLP (Manning et al., 2014), Keras[6]. Approximately 1/3 of the systems indicated that they used one or more of these ap-

[3] https://www.figure-eight.com/

[4] http://www.cs.cmu.edu/~ark/TweetNLP

[5] http://www.nltk.org/api/nltk.tokenize.html

[6] http://keras.io/preprocessing/text/

Tweet	A	B	C
@USER He is so generous with his offers.	NOT	—	—
IM FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE	OFF	UNT	—
@USER Fuk this fat cock sucker	OFF	TIN	IND
@USER Figures! What is wrong with these idiots? Thank God for @USER	OFF	TIN	GRP

Table 2: Four tweets from the OLID dataset, with their labels for each level of the annotation model.

proaches.

The results for each of the tasks are shown Table 3. Due to the large number of submissions, we only show the top 10 team F1-score results followed by result ranges for the rest of the teams. In addition to showing the results from the participating teams, we also include the models and baselines provided in Zampieri et. al. 2019. The models are a CNN, BiLSTM, and SVM. The baselines are choosing all predictions to be each class in the subtask (e.g. all offensive, and all not offensive for Subtask A). Table 4 shows all the teams that participated in the tasks along with their ranks in each task. These two tables can be used together to find the score/range for a particular team. The top three teams, by average rank for those that participated in all the sub-tasks were: *vradivchev_anikolov*, *NLPR@SRPOL*, and *NULI*. The following subsections describe the main results for each sub-task.

5.1 Sub-task A

Subtask A was the most popular task with 104 participating teams. 7/10 of the top teams used BERT (Devlin et al., 2018) with variations in parameters and pre-processing steps. The top performing team, *NULI*, used BERT-base-uncased with default-parameters but a max sentence length of 64 and trained with 2 epochs. The 82.9% F1 score of NULI is 1.4 points better than the next result, but the difference between the next top 5 systems, (ranked 2-6,) is less than one point (81.5%-80.6%) indicating that many of the top teams performed quite well. The top non-BERT model, *MIDAS* was in sixth place. They used an Ensemble approach of CNN and BLSTM+BGRU. They also used twitter word2vec embeddings (Godin et al., 2015) and token/hashtag normalization.

5.2 Sub-task B

76 teams participated in sub-task B. Most of these teams (71) also participated in sub-task A, but there were 5 new teams as well. In contrast to sub-task A, where BERT performed very well, 5

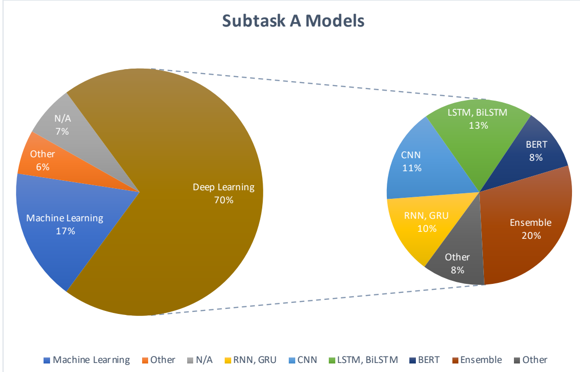

Figure 1: Pie chart showing the common models used in Subtask A. Deep Learning was the most popular. N/A indicates the systems that did not provide a description.

of the top 10 teams used an ensemble. Surprisingly, the best team, *jhan014*, which ranked 76 in subtask A, used a rule-based approach where they employed a keyword filter based on a Twitter language behavior list which included strings such as hash-tags, at signs, and they, etc... to achieve an F1 score of 75.5%. The second and third teams (*Amobee, HHU*) used ensembles of deep learning (including BERT) and machine learning. The best team in sub-task A did perform well at a rank of 4 (71.6) providing further indication that BERT does a good job at sub-task B as well.

5.3 Sub-task C

66 teams participated in the target identification task. The majority of these teams participated in sub-tasks A and B, but there were six new teams as well. As in sub-task B, ensembles were the most successful with 5/10 top teams having an ensemble of deep learning and machine learning methods. However, as in subtask A the best team, *vradivchev_anikolov*, used BERT after trying many other deep learning methods. They also used pre-processing and pretrained word embeddings (Glove). The second best team, *NLPR@SRPOL*, used an ensemble of OpenAI Finetune, deep learning models such as LSTM, transformer, and embeddings, and ma-

Subtask A		Subtask B		Subtask C	
Team Ranks	F1 Range	Team Ranks	F1 Range	Team Ranks	F1 Range
1	0.829	1	0.755	1	0.660
2	0.815	2	0.739	2	0.628
3	0.814	3	0.719	3	0.626
4	0.808	4	0.716	4	0.621
5	0.807	5	0.708	5	0.613
6	0.806	6	0.706	6	0.613
7	0.804	7	0.700	7	0.591
8	0.803	8	0.695	8	0.588
9	0.802	9	0.692	9	0.587
CNN	**0.800**	**CNN**	**0.690**	10	0.586
10	0.798	10	0.687	11-14	.571-.580
11-12	.793-.794	11-14	.680-.682	15-18	.560-.569
13-23	.782-.789	15-24	.660-.671	19-23	.547-.557
24-27	.772-.779	**BiLSTM**	**0.660**	24-29	.523-.535
28-31	.765-.768	25-29	.640-.655	30-33	.511-.515
32-40	.750-.759	**SVM**	**0.640**	34-40	.500-.509
BiLSTM	**0.750**	30-38	.600-.638	41-47	.480-.490
41-45	.740-.749	39-49	.553-.595	**CNN**	**0.470**
46-57	.730-.739	50-62	.500-.546	**BiLSTM**	**0.470**
58-63	.721-.729	**ALL TIN**	**0.470**	**SVM**	**0.450**
64-71	.713-.719	63-74	.418-.486	46-60	.401-.476
72-74	.704-.709	75	0.270	61-65	.249-.340
SVM	**0.690**	76	0.121	**All IND**	**0.210**
75-89	.619-.699	**All UNT**	**0.100**	**All GRP**	**0.180**
90-96	.500-.590			**ALL OTH**	**0.090**
97-103	.422-.492				
All NOT	**0.420**				
All OFF	**0.220**				
104	0.171				

Table 3: F1-Macro of the top 10 teams followed by the rest of the teams grouped in ranges for all three sub-tasks in terms of F1 Macro (%). Refer to Table 4 to see the team names associated with each rank. We also include the models and baselines provided in Zampieri et. al. (2019) in bold.

chine learning models such as SVM and Random forest.[7]

6 Conclusion

In this paper, we presented the results of SemEval-2016 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In OffensEval we used OLID (Zampieri et al., 2019), a dataset containing English tweets annotated with a hierarchical three-layer annotation model which considers 1) whether a message is offensive or not (sub-task A); 2) what is the type of the offensive message (sub-task B); and 3) what is the target of the offensive (sub-task C). OLID is publicly available to the research community.[8]

In total, nearly 800 teams signed up to participate in OffensEval and 115 of them submitted results across the three sub-tasks. In Section 5 we discussed the approaches used by the 115 teams in the shared task. We observed that both deep learning and traditional ML classifiers and classifier ensembles have been widely use and that most high-performing systems used state-of-the-art deep learning models, in particular BERT (Devlin et al., 2018). Our public dataset can continue

[7] In the camera-ready version of this report we will be including a Table with references to all system descriptions papers.

[8] `https://scholar.harvard.edu/malmasi/olid`

to be used to explore future advances in detecting offensive content and provide a a benchmark for evaluating different models. In the future, we plan to release additional content to address the class imbalance and small test size, particularly in subtasks B and C.

Acknowledgments

The research presented in this paper was partially supported by an ERAS fellowship awarded to Marcos Zampieri by the University of Wolverhampton.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In *Advances in Information Retrieval*, pages 693–696. Springer.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, pages 11–17.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of WWW*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of ACL*.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab $@$ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 146–153. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of TRAC*.

Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets Against Blacks. In *Proceedings of AAAI*.

Prasenjit Majumder, Thomas Mandl, et al. 2018. Filtering Aggression from the Multilingual Social Media Feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 199–207.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of RANLP*.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1 – 16.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Nithum Thain, Lucas Dixon, and Ellery Wulczyn. 2017. Wikipedia Talk Labels: Toxicity.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. *arXiv preprint arXiv:1705.09899*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval*.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from Bullying Traces in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

	Subtask Rank				Subtask Rank				Subtask Rank		
Team	A	B	C	Team	A	B	C	Team	A	B	C
NULI	**1**	4	18	resham	40	43	-	kroniker	79	71	-
vradivchev_anikolov	2	16	**1**	Xcosmos	41	47	29	aswathyprem	80	-	-
UM-IU@LING	3	76	27	jkolis	42	-	-	DeepAnalyzer	81	38	45
Embeddia	4	18	5	NIT_Agartala_NLP_Team	43	5	38	Code Lyoko	82	-	-
MIDAS	5	8	-	Stop PropagHate	44	-	-	rowantahseen	83	-	-
BNU-HKBU	6	62	39	KVETHZ	45	52	26	ramjib	84	-	-
SentiBERT	7	-	-	christoph.alt	46	14	36	OmerElshrief	85	-	-
NLPR@SRPOL	8	9	2	TECHSSN	47	22	16	desi	86	56	-
YNUWB	9	-	-	USF	48	32	62	Fermi	87	31	3
LTL-UDE	10	-	19	Ziv_Ben_David	49	64	33	mkannan	88	-	-
nlpUP	11	-	-	JCTICOL	50	63	-	mking	89	35	54
ConvAI	12	11	35	TüKaSt	51	23	50	ninab	90	69	-
Vadym	13	10	-	Gal_DD	52	66	25	dianalungu725	91	74	65
UHH-LT	14	21	13	HAD-Tübingen	53	59	61	Halamulki	92	-	-
CAMsterdam	15	19	20	Emad	54	-	-	SSN_NLP	93	65	64
YNU-HPCC	16	-	-	NLP@UIOWA	55	27	37	UTFPR	94	-	-
nishnik	17	-	-	INGEOTEC	56	15	12	rogersdepelle	95	-	-
Amobee	18	2	7	Duluth	57	39	44	Amimul Ihsan	96	-	-
himanisoni	19	46	11	Zeyad	58	34	34	supriyamandal	97	75	-
samsam	20	-	-	ShalomRochman	59	70	58	ramitpahwa	98	-	-
JU_ETCE_17_21	21	50	47	stefaniehegele	60	-	-	ASE - CSE	99	33	32
DA-LD-Hildesheim	22	28	21	NLP-CIC	61	48	46	kripo	100	-	-
YNU-HPCC	23	12	4	Elyash	62	67	40	garain	101	44	63
ChenXiuling	24	-	28	KMI_Coling	63	45	53	NAYEL	102	-	-
Ghmerti	25	29	-	RUG_OffenseEval	64	-	-	magnito60	103	-	-
safina	26	-	-	jaypee1996	65	41	-	AyushS	104	36	48
Arjun Roy	27	17	-	orabia	66	55	8	UBC_NLP	-	6	9
CN-HIT-MI.T	28	30	22	v.gambhir15	67	58	60	bhanodaig	-	57	-
LaSTUS/TALN	29	20	15	kerner-jct.ac.il	68	68	42	Panaetius	-	60	-
HHU	30	3	-	SINAI	69	-	-	eruppert	-	61	-
na14	31	26	10	apalmer	70	13	55	Macporal	-	72	-
NRC	32	37	24	ayman	71	53	57	NoOffense	-	-	6
NLP	33	54	52	Geetika	72	24	-	HHU	-	-	14
JTML	34	-	-	Taha	73	51	59	quanzhi	-	-	17
Arup-Baruah	35	25	31	justhalf	74	-	-	TUVD	-	-	23
UVA_Wahoos	36	42	-	Pardeep	75	7	41	mmfouad	-	-	51
NLP@UniBuc	37	73	49	**jhan014**	76	**1**	30	balangheorghe	-	-	56
NTUA-ISLab	38	40	43	liuxy94	77	-	-				
Rohit	39	49	-	ngre1989	78	-	-				

Table 4: All the teams participating in Offenseval 2019 with their ranks for each sub-task. - indicates no participation. Refer to Table 3 to see the scores based on a team's rank. The top team for each task is in **bold**, and the second place team is underlined. ASE - CSE stands for Amrita School of Engineering - CSE and BNU-HBKU stands for BNU-HKBU UIC NLP Team 2.