Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages

Sandip Modha[1,4][0000−0003−2427−2433], Thomas Mandl[2,3][0000−0002−8398−9699],
Prasenjit Majumder[3], and Daksh Patel[4]

[1] DA-IICT,Gandhinagar,India sjmodha@gmail.com
[2] University of Hildesheim, Germany mandl@uni-hildesheim.de
[3] DA-IICT,Gandhinagar,India prasenjit.majumder@gmail.com
[4] LDRP-ITR,Gandhinagar,India dakshpatel68@gmail.com

Abstract. The identification of Hate Speech in Social Media has received much attention in research recently. There is a particular demand for research for languages other than English. The first edition of the HASOC track creates resources for Hate Speech Identification in Hindi, German, and English. Three datasets were developed from Twitter, and Facebook and made available. HASOC intends to stimulate research and development for Hate Speech classification for different languages. The datasets allow the development and testing of supervised machine learning systems. Binary classification and more fine-grained sub-classes were offered in 3 sub tasks. For all sub-tasks, 321 experiments were submitted. For the classification task, models based on deep learning methods have proved to be adequate. The approaches used most often were Long-Short-Term memory (LSTM) networks with distributed word representation of the text. The performance of the best system for identification of Hate Speech for English, Hindi, and German was a Marco-F1 score of 0.78, 0.81, and 0.61, respectively. This overview provides details insights and analyzes the results.

Keywords: Hate Speech · Text Classification · Evaluation· Deep Learning.

1 Introduction

The large fraction of Hate Speech and other offensive and objectionable content online poses a huge challenge to societies. Offensive language such as insulting, hurtful, derogatory, or obscene content directed from one person to another person and open for others undermines objective discussions. There is a growing need for research on the classification of Hate Speech into different categories of offensive content on different platforms of social media without human assistance.

In October 2019, the European Court of Justice decided that platforms need to take down content worldwide even after national decisions. In a particular case, the EU court debated defamatory posts on Facebook. Even posts similar in tone need to be addressed and the ruling explicitly mentions automatic systems. This shows that automatic systems are of high social relevance. Recently, also the founder of Facebook proposed ideas for the regulation of the Internet. He demanded standards and baselines for the definition of harmful content. Such clear definitions have not been provided and are unlikely to be developed in the near future. This makes research and annotated corpora even more necessary.

The identification of Hate Speech within a collection or a stream of tweets is a challenging task because systems cannot rely on the text content. Based on content, text classification systems have been successful. However, Hate text might have many issues. Hate often has no clear signal words, and word lists, as in sentiment analysis, are expected to work less well.

In order to contribute to this research, w this overview paper presents the 1st edition of HASOC Hate Speech and Offensive Content Identification in Indo-European Languages, namely: German, English, and Hindi. The dataset for all three languages was created from Twitter and Facebook. HASOC consists of three tasks, a coarse-grained binary classification task, and two fine-grained multi-class classifications. Of course, freedom of speech needs to be guaranteed in democratic societies for future development. Nevertheless, the offensive text which hurts others' sentiments needs to be restricted. As there is such an increase in the usage of abuse on many internet platforms, technological support for the recognition of such posts is necessary. The use of supervised learning with the annotated dataset is a key strategy for advancing such systems. There has been significant work in several languages in particular for English. However, there is a lack of research on this recent and relevant topic for most other languages. This track intends to develop data and evaluation resources for several languages. The objectives are to stimulate research for these languages and to find out the quality of hate speech detection technology in other languages.

The HASOC dataset provides several thousands labeled social media posts for each language. The entire dataset was annotated and checked by the organizers of the track. The annotation architecture is designed to create data for 3 different sub tasks.

1. SUB-TASK A: classification of Hate Speech (HOF) and non-offensive content.
2. SUB-TASK B: If the post is HOF, sub-task B is used to identify the type of hate.
3. SUB-TASK C: it decides the target of the post.

Hate Speech detection is of great significance and attracting many researchers. Recent overview papers provide a good introduction to the scientific issues that are involved in Hate Speech identification [12,36].

https://www.nytimes.com/2019/10/03/technology/facebook-europe.html

https://www.faz.net/aktuell/wirtschaft/diginomics/facebook-ceo-zuckerberg-ideas-to-regulate-the-internet-16116032.html

2 Related Forum and Dataset

Collections are an important asset for any supervised classification methods. For Hate Speech, several previous initiatives have created corpora that have been used for research. There has been significant work in several languages, in particular for English. However, for other languages, such as Hindi standard datasets are not available and HASOC is an attempt to create the labeled dataset for such low resource language. HASOC is primarily inspired by two previous evaluation forums, GermEval [44], and OffensEval [47], and tries to leverage the synergies of these initiatives.

Data sampling is a paramount task for any data challenges competition. Some of the corpora focuses in specific on certain targets, like immigrants, women (HateEval) [5]or racism (e.g. [39]). Others focus on Hate Speech in general (e.g. HaSpeeDe [7]) or other unacceptable text types. A recent trend is to introduce a more fine-grained classification. Some data challenges require detailed analysis for the hateful comments, like detection of the target (HateEval and OffensEval) or the type of Hate Speech (GermEval). Others focus on the severity of the comment (Kaggle Toxic [1]). A recent and very interesting collection is CONAN. It offers Hate Speech and the reactions to it [9]. This could open opportunities for detecting Hate Speech by analyzing it jointly with the following posts. Table 1 summarize standard Hate speech dataset available at various forum.

There is a huge demand for many languages other than English. HASOC is the first shared task which developed a resource for three languages together and which encourages multilingual research.

3 Task Description

HASOC and most other collections provide the text of a post and require systems to detect hateful content. No context or meta-data like time related features or the network of the actors are given which might make these tasks somewhat unrealistic. Platforms can obviously use all meta-data of a post and a user. However, the distribution of such data poses legal issues. The following tasks have been proposed in HASOC 2019:

Sub-task A : Sub-task A focuses on Hate speech and Offensive language identification and is offered for English, German, Hindi. Sub-task A is coarse-grained binary classification in which participating system are required to classify tweets into two class, namely: Hate and Offensive (HOF) and Non- Hate and offensive.

1. (NOT) Non Hate-Offensive - This post does not contain any Hate speech, offensive content.
2. (HOF) Hate and Offensive - This post contains Hate, offensive, and profane content.

During our annotation, we labeled posts as HOF in case they contained any form of non-acceptable language such as hate speech, aggression, profanity; otherwise they were labeled as NOT.

Table 1. Recent Collections for Research on Offensive Content Detection

Dataset	**Source**	**Language**	**# labelled posts**	**Task**	**Metric and Best Besult**
GermEval Task2 2019 [38]	Twitter	German	4000	3 levels, Hate, type, implicit/ explicit	Macro F1 0.76
OffensEval at SemEval [45]	Twitter	English	13200	3 levels, Hate, targeted and target type	F1 score 0.83
HateEval at SemEval [5]	Twitter	Spanish, English	19000	Hate, aggression, target	Macro F1 0.65 Engl.
Kaggle Toxic [1]	Wikipedia comments	mostly English	240000	5-class	Column-wise AUC 0.98
Racism [18]	Twitter	English	24000	Binary, Racism	Accuracy 0.76
TRAC COLING [30]	Facebook, Twitter	English, Hindi	15000 each language	3 classes, overtly or covertly aggressive	weighted F1-score 0.64
Arabic Social Media [22]	Twitter,	Arabic	1100 tweets, 32000 comments	Obscene, inappropriate	F1 around 0.60
Racism Detection in Social Media [40]	Belgian social media sites	Dutch	5400	Binary, Racism	F1 score 0.46
Offensive Language [11]	Twitter	English	14500	Binary, Hate	F1 score 0.90

Sub-task B : Sub-task B represents a fine-grained classification. Hate-speech and offensive posts from the sub-task A are further classified into three categories.

1. (HATE) Hate speech: Posts contain Hate speech content.
2. (OFFN) Offensive: Posts contain offensive content.
3. (PRFN) Profane: These posts contain profane words

HATE SPEECH : Describing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. all poor people are stupid). Hateful comment toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar.

OFFENSIVE : Posts which are degrading, dehumanizing, insulting an individual, threatening with violent acts are categorized into this category.

PROFANITY : Unacceptable language in the absence of insults and abuse. This typically concerns the usage of swearwords (Scheiße, Fuck etc.) and cursing (Hell! Verdammt! etc.). Such posts are categorized into this category. As expected, most posts are in the category NOT, some are HATE and the other two categories are less frequent. Dubious cases which are difficult to decide even for humans, were left out.

Sub-task C (only for English and Hindi) : Sub-task C considers the type of offense. Only posts labeled as HOF in sub-task A are included in sub-task C. The two categories in sub-task C are the following:

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

4 Data Set and Collection

The following sections explain how the data set was created and enriched by annotations. First, the authors searched with heuristics for typical Hate Speech in online fora. They identified topics for which many hate posts can be expected. Different hashtags and keywords were used for all three languages. For some of the found posts, the id of the author was recorded. For a number of such users, the timeline was collected. Based on tweets found, we crawled the last posts of the authors to increase variety. The systems are less likely to classify individual textual style when they have a rich set of posts from an author. This procedure was intended to decrease bias and was inspired by GermEval [43].

The HASOC dataset was subsequently sampled from Twitter and partially from Facebook for all the three languages. The Twitter API gives a large number

Table 2. Collection and Class Distribution for Training Set

Lang.	NOT	HOF	HATE	OFFN	PRFN	Total
English	3591	2261	1143	667	451	5852
Hindi	2196	2469	556	676	1237	4665
German	3412	407	111	210	86	3819

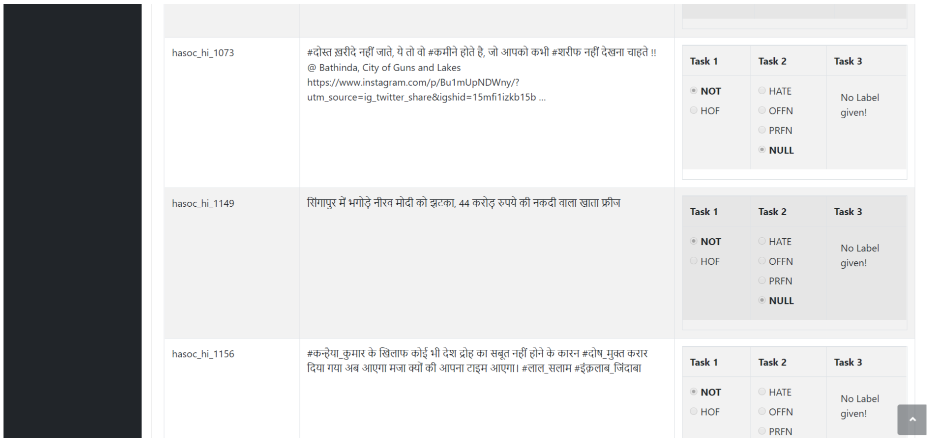
Table 3. Collection and Class Distribution for test Dataset

Lang.	NOT	HOF	HATE	OFFN	PRFN	Total
English	865	288	124	71	93	1153
Hindi	713	605	190	197	218	1318
German	714	136	41	77	18	850

Table 4. Example Tweets for all Classes

Classes	Sample tweet from the class
NOT	4 matches were can't play due to rain and many more will be not played fir the same reason . Conclusion this world cup is no more world cup. #ShameOnICC #RainCup
HATE	Are Muslims, in general a nuisance to be tolerated by the rest of the world ? #SaveBengal #DoctorsFightBack #DoctorsStrike #MamtaBanerjee
HATE	#TerroristNationPakistan 90% Pakistanis wants war with India and 10% said war should not be. And Those 10% belongs to Pakistans Armed Forces #TerroristNationPakistan
OFFN	#Just a daily reminder to @realDonaldTrump that he is a National Disgrace. #TraitorTrump #TrumpIsADisgrace #TrumpIsATraitor
PRFN	@cizzacampbell Didn't realise you were an expert #dickhead
PRFN UNT	Who voted for a no-deal? Tell me, who the fuck voted for a no deal? The way I see it, the referendum was a corrupt vote between remain and leave. Not remain, leave, deal, no deal. Nobody voted for no deal!!
OFFN TIN	@realDonaldTrump Will it be worse than killing children? Worse than selling your country to the Russians? Worse than saying you love a ruthless dictator? Probably not. #TrumpIsATraitor

of recent tweets which resulted in an unbiased dataset. Thus the tweets were acquired using hashtags and keywords that contained offensive content. The collection was provided to participants without metadata. We have developed Twitter and Facebook plugins to fetch the posts without using the API. The size of the data corpus is shown in tables 2 and 3.



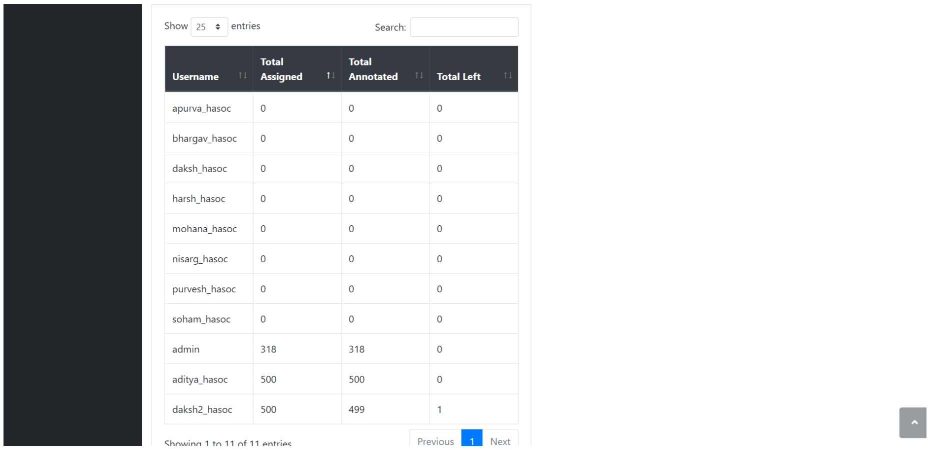
		Task 1	Task 2	Task 3
hasoc_hi_1073	#दोस्त ख़रीदे नहीं जाते, ये तो वो #कमीने होते है, जो आपको कभी #शरीफ नहीं देखना चाहते !! @ Bathinda, City of Guns and Lakes https://www.instagram.com/p/Bu1mUpNDWny/?utm_source=ig_twitter_share&igshid=15mfi1izkb15b ...	● NOT ○ HOF	○ HATE ○ OFFN ○ PRFN ● NULL	No Label given!
hasoc_hi_1149	सिंगापुर में भगोड़े नीरव मोदी को झटका, 44 करोड़ रुपये की नकदी वाला खाता फ्रीज	● NOT ○ HOF	○ HATE ○ OFFN ○ PRFN ● NULL	No Label given!
hasoc_hi_1156	#कन्हैया_कुमार के खिलाफ कोई भी देश द्रोह का सबूत नहीं होने के कारन #दोष_मुक्त करार दिया गया अब आएगा मजा क्यों की आपना टाइम आएगा। #लाल_सलाम #इंक़लाब_जिंदाबा	● NOT ○ HOF	○ HATE ○ OFFN ○ PRFN	No Label given!

Fig. 1. Screenshot of online Annotation System.

During the labeling process, several juniors for each language engaged with an online system to judge the tweets. The system can be seen in figure 1 and figure 2. They were given short guidelines that contained the information as mentioned in section 3.1. The process is highly subjective, and even after discussions of questionable often no agreement could be reached. This lies in the nature of Hate Speech.

As pointed out in the study by Ross et al. [32], not even with providing written guidelines can improve the agreement. Consequently, and to be sure that people can see them on one page, we tried to keep the guidelines short. The guidelines for HASOC are listed in the annex. A study by Salminen et al. [35] showed that the dubious and questionable cases led to much more disagreement than clear cases with obvious Hate Speech characters. Jhaver et al. [14] and colleagues interviewed both the receivers and the senders of some posts which were considered to be aggressive. They revealed that the senders often did not agree with the judgment of readers. Among other arguments, they brought forward that some messages were regarded as hateful because people did not want to be confronted with the arguments. Again, this study shows that there is a great deal of subjectivity involved and that also context matters.

The difficulties during assessment in HASOC were often related to the use of language registers like youth talk and irony or indirectness which might not



Show 25 entries

Search:

Username	Total Assigned	Total Annotated	Total Left
apurva_hasoc	0	0	0
bhargav_hasoc	0	0	0
daksh_hasoc	0	0	0
harsh_hasoc	0	0	0
mohana_hasoc	0	0	0
nisarg_hasoc	0	0	0
purvesh_hasoc	0	0	0
soham_hasoc	0	0	0
admin	318	318	0
aditya_hasoc	500	500	0
daksh2_hasoc	500	499	1

Showing 1 to 11 of 11 entries

Fig. 2. Screenshot of Statistics Module of online Annotation System

be understood by all readers. A more detailed analysis of the issues encountered during the HASOC annotation for German has been carried out[42]. u

The overlap between annotators for task A for English, Hindi, and German for a subset to tweets and posts annotated twice was 89%, 91%, and 32%, respectively. Further statistical details of the annotation process can be seen in table 5. The effects of such disagreement need to be analyzed in the future.

Table 5. Interrater Statistics on HASOC Multilingual Datasets

Task	No. of Posts annotated twice/Total Posts	Percentage of Posts annotated twice	No. of Posts with same annotation	Interrater Agreement
English sub-task A	6246/7005	89%	4838	77.46%
English sub-task B	6246/7005	89%	4311	69.02%
English sub-task C	6246/7005	89%	4669	74.75%
Hindi sub-task A	5440/5983	91%	4281	78.69%
Hindi sub-task B	5440/5983	91%	3421	62.89%
Hindi sub-task C	5440/5983	91%	3488	64.12%
German sub-task A	1483/4669	32%	1305	88%
German sub-task B	1483/4669	32%	1283	86.51%

The values show that the labeling task is hard overall. The second sub-task can only be solved with a lower quality. For the sub-task C, the quality does not drop much of is even higher than that of sub-task B .

We also calculated the κ (Kappa) coefficient due to the high imbalance of the data sets. Using the scikit-Learn package, the inter annotator agreement for the first two annotators for a tweet was determined. Table 6 shows values of κ in sub-task A for all three languages

Table 6. κ statistics

Language	Sub-task A
English	0.36
Hindi	0.59
German	0.43

The degree of disagreement might also result from the topics present in the collection[46]. The issues and the level of disagreement need to be analyzed in the future.

5 Evaluation Metrics

The metrics for classification should combine both recall and precision. The F1-score has many variants like weighted F1, Macro-F1 or micro-F1. For multi-class classification, the distribution of class labels is often unbalanced. The weighted F1-score calculates the F1 score for each class independently. When it adds them, it uses a weight based on the number of true labels of each class. Therefore, it gives a bias for the majority class. The 'macro' calculates the F1 separately for each class but does not use weights for the aggregation. This results in a stronger penalization when a system does not perform well for the minority classes. Choice of the variant of F1-measure depends on the objective of the tasks and the distribution of label in the dataset. Hate Speech related classification problems suffer from class imbalance. Therefore, the macro F1 is the natural choice for the evaluation.

6 Results

Overall, 103 registrations were submitted for the track. 37 teams submitted runs and 25 teams have submitted papers. 321 runs were submitted by 37 teams in all the sub-tasks.

Table 7. Number of Experiments Submitted

Languages	sub-task A	sub-task B	sub-task C
English	79	50	45
Hindi	37	31	25
German	28	26	n.a.

The following sections show the sub-tasks of HASOC. The approaches of all teams are briefly summarized in the annex of this paper. For details on the technical implementation, the reader is referred to the descriptions of the participating teams in this volume.

6.1 English Dataset

In the English language, Total 174 runs were submitted across 3 sub-tasks. The YNU_wb team [6] used an LSTM approach with ordered neurons and applied an attention mechanism. The absolute differences between the top runs are rather small. Table 8 presents the results of the top 10 teams of the English sub-task A.

Table 8. Best Runs for English Sub-Task A

Standing	Team name	Run_no	Marco F1	Weighted F1
1	YNU_wb [6]	2	0.7882	0.8395
2	YNU_wb [6]	3	0.772	0.8237
3	BRUMS [29]	2	0.7694	0.838
4	YNU_wb [6]	1	0.7682	0.8175
5	vito [25]	2	0.7568	0.8182
6	vito [25]	3	0.7471	0.8071
7	3Idiots [21]	2	0.7465	0.8012
8	IIITG-ADBU [3]	1	0.7462	0.8064
9	QMUL-NLP [15]	1	0.7431	0.8164
10	RALIGRAPH [19]	3	0.7409	0.7876
11	3Idiots [21]	2	0.8004	0.801
12	QutNocturnal [4]	2	0.8002	0.8001
13	LSV-UdS [10]	3	0.7996	0.7995
14	IIITG-ADBU [3]	3	0.7985	0.7986
15	NITK-IT_NLP	2	0.7889	0.7888
16	LSV-UdS [10]	2	0.7837	0.784
17	Kirti Kumari [17]	2	0.7827	0.7826
18	HateMonitors [34]	1	0.7754	0.7759
19	DEEP [24]	1	0.7594	0.7592
20	FalsePostive [16]	2	0.756	0.756

The plot of the performance of all systems in Figure 3 shows that the Median of the runs lies quite close to the top performance.

Despite the similar performance of many teams, the recall-precision graph in figure 4 shows that there are considerable differences between the systems which the F1 measures do not reveal.

The overall F1 measures for sub-task B and C are much lower than for sub-task A. Table 9 and 10 shows the results of these tasks. The best performing team [21] for sub-task B and sub-task C used the relatively new BERT model for classification. This shows that it performed well for both sub-task A with more training samples as well as for sub-task B with much fewer training instances.

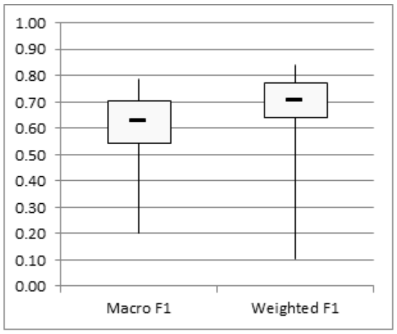


Fig. 3. Box-plot of the performance of all runs for English sub-task A.

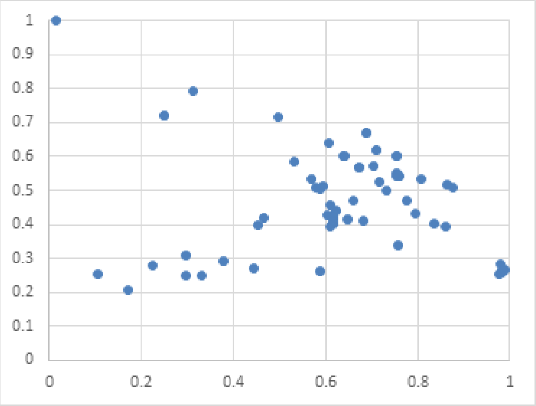


Fig. 4. Recall-Precision Graph of all Runs for the English sub-task A

Table 9. Best Runs for sub-task B English Dataset

Standing	Team name	Run_no	Marco F1	Weighted F1
1	3Idiots [21]	3	0.5446	0.7277
2	3Idiots[21]	2	0.537	0.698
3	3Idiots [21]	1	0.5175	0.701
4	VITO [25]	3	0.5064	0.7514
5	VITO [25]	2	0.5051	0.7595
6	RALIGRAPH [19]	2	0.4789	0.7218
7	RALIGRAPH [19]	1	0.4777	0.7147
8	RALIGRAPH [19]	3	0.4732	0.6911
9	LSV-UdS [19]	2	0.4658	0.4948
10	QutNocturnal [4]	1	0.4501	0.6813

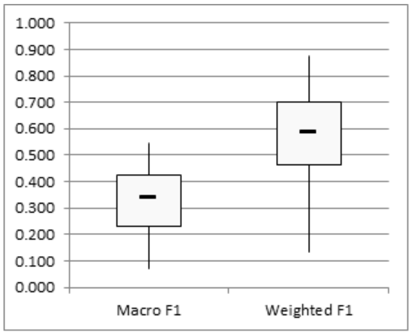


Fig. 5. Box-plot of the performance of all runs for English sub-task B

Table 10. Best Runs for sub-task C English Dataset

Standing	Team name	Run_no	Marco F1	Weighted F1
1	3Idiots [21]	3	0.5111	0.7563
2	3Idiots[21]	1	0.5002	0.753
3	VITO [25]	4	0.494	0.7735
4	RALIGRAPHv[19]	2	0.4907	0.7719
5	VITO[25]	2	0.4879	0.784
6	3Idiots [21]	2	0.4765	0.7639
7	RALIGRAPH [19]	3	0.4758	0.7302
8	HateMonitors [34]	1	0.4698	0.7057
9	RALIGRAPH [19]	1	0.4639	0.7265
10	IRLAB@IITBHU [2]	2	0.4578	0.7704

The performance for task C shows that the weighted F1 values are very close together and that run number 10 has even a higher values than run number 1. The careful selection of metrics is crucial. The boxplots in figure 5, and 6 show that the Median lies again close to the top performing run for sub-tasks B and C.

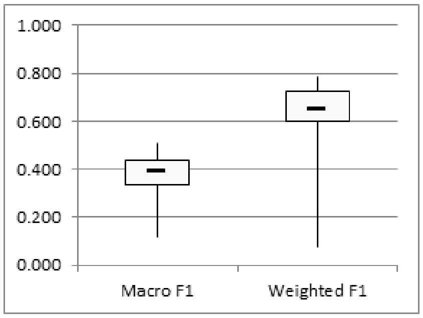


Fig. 6. Box-plot of the performance of all runs for English sub-task c

6.2 Hindi Dataset

In the Hindi language, total 93 runs were submitted across 3 sub-tasks. The QutNocturnal team [4] used a CNN base approach with Word2vec embedding. The absolute differences between the top runs are rather small. Table 11 presents results of the top team of Hindi sub-task A The absolute values for Hindi sub-task A are comparable to the English sub-task and the top-performing systems are again close to each other.

Table 11. Best Runs for sub-task A Hindi Dataset

Standing	Team name	Run_no	Marco F1	Weighted F1
1	QutNocturnal [4]	1	0.8149	0.8202
2	LGI2P [13]	2	0.8111	0.8116
3	3Idiots [21]	3	0.8108	0.8141
4	IIITG-ADBU [3]	2	0.8105	0.8108
5	IIITG-ADBU [3]	1	0.8098	0.81
6	LGI2P [13]	1	0.8076	0.8088
7	A3-108 [23]	2	0.8032	0.8032
8	brum [29]	1	0.8025	0.8024
9	A3-108 [23]	3	0.8024	0.8024
10	3Idiots [21]	1	0.8018	0.8025

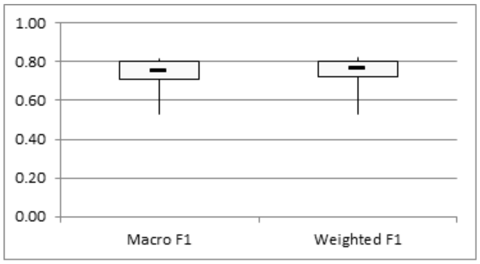


Fig. 7. Box-plot of the performance of all runs for Hindi sub-task A

Table 12. Best Runs for sub-task B Hindi Dataset

Standing	Team name	Run_no	Marco F1	Weighted F1
1	3Idiots	3	0.5812	0.7147
2	LSV-UdS [10]	3	0.5779	0.6358
3	LSV-UdS [10]	2	0.5692	0.6386
4	LGI2P [13]	3	0.5617	0.674
5	QutNocturnal [4]	1	0.561	0.6551
6	3Idiots [21]	2	0.5534	0.6755
7	3Idiots [21]	1	0.5527	0.6875
8	LSV-UdS [10]	1	0.5392	0.5504
9	A3-108 [23]	2	0.5253	0.756
10	A3-108 [23]	3	0.5113	0.7514

Table 12 and 13 presents result of Hindi sub-task B and C.The overall values for sub-task B and C for Hindi are again comparable to the values for English. Figure 7, 8,9 shows the overall performance of all teams for all Hindi sub-tasks.

6.3 German Dataset

In the German language, total 54 runs were submitted across 2 sub-tasks and only, the first two sub-tasks were possible. The Macro F1 score is lower than for the other two languages. For sub-task A, the best team used BERT sentence embedding and the multilingual sentence embedding LASER. Table 14 and 15 present result of sub-task A and B.

The LSV team [10] performed second and first for sub-task B. They apply the BERT model and use additional corpora for similar tasks. Boxplots of the performance of all the participants team are shown in figure 10 and 11.

7 Approaches

The top performance for the sub-task A for English and and Hindi three languages is delivered by systems based on Deep neural models. Even new architectures for which little experience is available like BERT have been applied

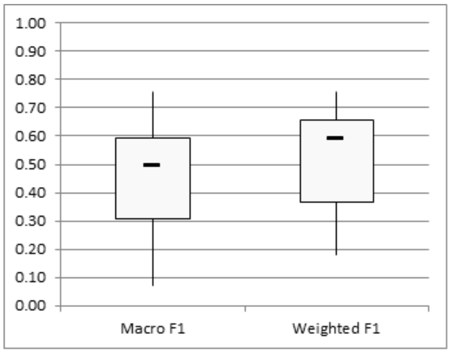


Fig. 8. Box-plot of the performance of all runs for Hindi sub-task B

Table 13. Best Runs for sub-task c Hindi Dataset

Standing	Team name	Run_no	Marco F1	Weighted F1
1	A3-108 [23]	3	0.5754	0.7361
2	3Idiots [21]	1	0.565	0.7265
3	A3-108 [23]	2	0.5559	0.7447
4	3Idiots [21]	3	0.5503	0.7583
5	3Idiots [21]	2	0.5492	0.7484
6	DEEP [24]	3	0.5238	0.6803
7	DEEP [24]	1	0.5172	0.6967
8	QutNocturnal[4]	1	0.5165	0.7429
9	KMI-Panlingua [31]	1	0.497	0.6499
10	KMI-Panlingua [31]	2	0.497	0.6499

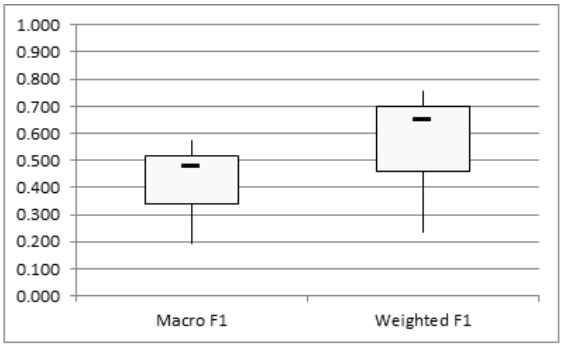


Fig. 9. Box-plot of the performance of all runs for Hindi sub-task C

Table 14. Best Runs for Sub-task A German Dataset

Standing	Team name	Run_no	Marco F1	Weighted F1
1	HateMonitors [34]	1	0.6162	0.7915
2	LSV-UdS [10]	1	0.6064	0.7997
3	LSV-UdS [10]	2	0.5948	0.7799
4	3Idiots [21]	1	0.5774	0.7887
5	NITK-IT_NLP	1	0.5739	0.6796
6	DLRG [28]	2	0.5519	0.7566
7	CS	1	0.5506	0.7131
8	BRUMS [29]	1	0.5464	0.787
9	DLRG [28]	1	0.5458	0.7816
10	LSV-UdS [10]	2	0.5399	0.7762

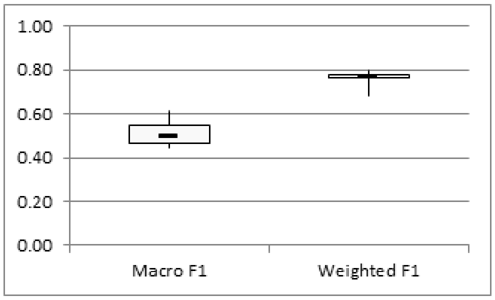


Fig. 10. Box-plot of the performance of all runs for German sub-task A

Table 15. Best Runs for Sub-task B German Dataset

Standing	Team name	Run_no	Marco F1	Weighted F1
1	LSV-UdS [10]	1	0.3468	0.7749
2	LSV-UdS [10]	3	0.2785	0.5829
3	HateMonitors [34]	1	0.2769	0.7537
4	3Idiots [21]	2	0.2758	0.7779
5	Cs	1	0.274	0.757
6	3Idiots [21]	3	0.2736	0.7729
7	FalsePostive [16]	3	0.268	0.7458
8	FalsePostive [16]	2	0.2619	0.7436
9	FalsePostive [16]	1	0.2608	0.7536
10	LSV-UdS [10]	2	0.2558	0.7545

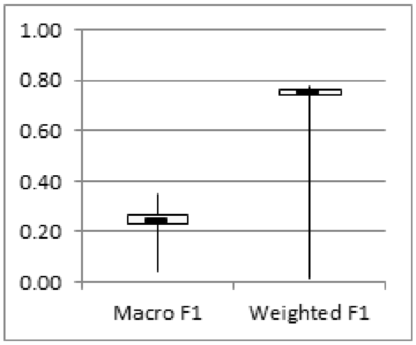


Fig. 11. Box-plot of the performance of all runs for German sub-task B

with great success. There is even true for sub-task B for German where only few training examples were available. There needs to be considered that most systems applied a Deep Learning system (see annex B). However, for Hindi the top performance comes from a traditional machine learning system. Even for the other two languages, we can observe that some of the few non-Deep Learning systems lead to a performance quite close to the top performance. For example, Team A3-108 [23] reaches a result close to the top performance for the Hindi sub-task B. Also the run IRLAB@IITBHU [2] achieves a higher weighted F1 value than the top run for sub-task B for English. It seems that the size of HASOC is small enough that traditional approaches can still prevail. There might not be enough data to train Deep architectures with many parameters. Future improvement for such systems might lie in the intelligent use of external resources. Participants were allowed to use external resources and other datasets for this task. For German, this seems to have boosted the top performing team LSV-UdS for the sub-task B for which only few training examples were available.

Several teams have adopted an open code policy and published their code in Github repositories. This policy allows repeat-ability and reproducibility of the experiments.

8 Performance Analysis

Some of the participants have conducted an interesting analysis in order to explore the behavior of their systems. We tried to explore the performance of all systems on each tweet. We ranked the tweets for sub-task A in English based on the number of systems that classified them. The following figure shows the distribution of the values.

We can observe that only 30% of the systems agree a post is an offensive (class HOF) considering the Median. On the other hand, 70% of the systems vote for NOT in the Median for the class NOT. However, the distributions are quite scattered. This shows that for the systems there seem to be no clear and obvious

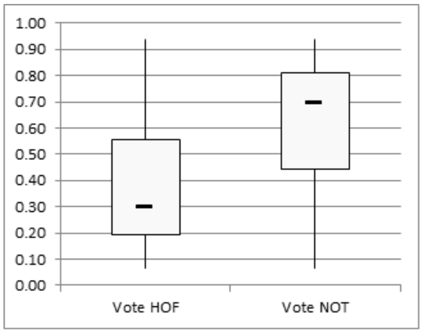


Fig. 12. Box-plot of the percentage of systems votes for tweets for English sub-task A

cases. Considering the analysis of Salminen et al. in which humans agreed much on obvious Hate Speech tweets [35], there seems to be less agreement by systems. As a consequence, voting approaches might not work well. Another consequence could be that it is hard to explain and understand the decision of a classifier in this domain. This may lead to a lack of ability to explain decisions and a lack of transparency. This can result in a low degree of acceptance in society. More analysis of the results is necessary for the future.

9 Conclusion and Outlook

The submissions for HASOC have shown that deep learning representations seem to be the state of the art approach for Hate Speech classification. After analyzing the results, the best method for Hate speech classification is dependent on the corpus language, classification granularities, and distribution of each class-labels. In other words balance, an unbalanced training dataset might affect the performance of the classification system. In the long run, the HASOC track aims at supporting researchers to develop robust technology which can cope with multilingual data and to develop transfer learning approaches that can exploit learning data across languages. For future editions, we envision the integration of further languages. The potential bias in the data collection needs to be analyzed and monitored [43].

10 Acknowledgements

We thank all participants for their submissions and the work involved. We thank all the jurors who labeled the tweets in a short period of time. We also thank the FIRE organizers for their support in organizing the track.

References

1. Kaggle (2017): Toxic comment classification challenge: Identify and classify toxic online comments, https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
2. Anita Saroj, R.K.M., Pal, S.: Irlab@iitbhu at hasoc 2019 2019:traditional machine learning for hate speech and offensive content identification. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
3. Baruah, A., Barbhuiya, F., Dey, K.: IIITG-ADBU at HASOC 2019: Automated Hate Speech and Offensive Content Detection in English and Code-Mixed Hindi Text. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
4. Bashar, M.A., Nayak, R.: QutNocturnalHASOC'19: CNN for Hate Speech and Offensive Content Identification in Hindi Language. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
5. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M.: Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 54–63 (2019)
6. Bin Wang, Yunxia Ding, S.L., Zhou, X.: YNU_Wb at HASOC 2019: Ordered Neurons LSTM with Attention for Identifying Hate Speech and Offensive Language. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
7. Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., Maurizio, T.: Overview of the evalita 2018 hate speech detection task. In: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. vol. 2263, pp. 1–9. CEUR (2018)
8. Casavantes, M., López, R., González, L.C., Montes-y Gómez, M.: UACh-INAOE at HASOC 2019: Detecting Aggressive Tweets by Incorporating Authors' Traits as Descriptors. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
9. Chung, Y.L., Kuzmenko, E., Tekiroglu, S.S., Guerini, M.: Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. arXiv preprint arXiv:1910.03270 (2019)
10. Dana Ruiter, M.A.R., Klakow, D.: LSVUdS at HASOC 2019: The Problem of Defining Hate? In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
11. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated Hate Speech Detection and the Problem of Offensive Language. In: Proceedings of ICWSM (2017)
12. Fortuna, P., Nunes, S.: A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys (CSUR) **51**(4), 85 (2018)
13. Jean-Christophe Mensonides, Pierre-Antoine Jean, A.T., Harispe, S.: IMT Mines Ales at HASOC 2019: Automatic Hate Speech detection. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
14. Jhaver, S., Ghoshal, S., Bruckman, A., Gilbert, E.: Online harassment and content moderation: The case of blocklists. ACM Transactions on Computer-Human Interaction (TOCHI) **25**(2), 12 (2018)

15. Jiang, A.: QMUL-NLP at HASOC 2019: Offensive Content Detection and Classification in Social Media. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
16. Kaushik Amar Das, F.A.B.: FalsePostive at HASOC 2019: Transfer-Learning for Detection and Classification of Hate Speech. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
17. Kirti Kumari, J.P.S.: AI_ML_NIT Patna at HASOC 2019: Deep Learning Approach for Identification of Abusive Content. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
18. Kwok, I., Wang, Y.: Locate the hate: Detecting Tweets Against Blacks. In: Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)
19. Lu, Z., Nie, J.Y.: RALIGRAPH at HASOC 2019: VGCN-BERT: Augmenting BERT with Graph Embedding for Offensive Language Detection. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
20. Mishra, A., Pal, S.: IIT Varanasi at HASOC 2019 : Hate Speech and Offensive Content Identification in Indo-European Languages . In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
21. Mishra, S., Mishra, S.: 3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
22. Mubarak, H., Darwish, K., Magdy, W.: Abusive language detection on arabic social media. In: Proceedings of the First Workshop on Abusive Language Online. pp. 52–56 (2017)
23. Mujadia, V., Mishra, P., Sharma, D.M.: IIIT-Hyderabad at HASOC 2019: Hate Speech Detection. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
24. Nayel, H.A., Shashirekha, H.L.: DEEP at HASOC2019 : A Machine Learning Framework for Hate Speech and Offensive Language Detection. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
25. Nina-Alcocer, V.: Vito at HASOC 2019: Detecting Hate Speech and Offensive Content through Ensembles. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
26. Parikh, A., Desai, H., Bisht, A.S.: DA_Master at HASOC 2019: Identification of Hate Speech using Machine Learning and Deep Learning approaches for social media post. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
27. Pedro Alonso, R.S., Kovács, G.: TheNorth at HASOC 2019: Hate Speech Detection in Social Media Data. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
28. R. Rajalakshmi, Y.R.: DLRG@HASOC 2019: An Enhanced Ensemble Classifier for Hate and Offensive Content Identification . In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
29. Ranasinghe, T., Zampieri, M., Hettiarachchi, H.: BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)

30. Ritesh, K., N., R.A., Akshit, B., MaheshwariTushar: Aggression-annotated corpus of hindi-english code-mixed data. In: Proceedings of the 11th Language Resources and Evaluation Conference (LREC). pp. 1–11. Miyazaki, Japan (2018)
31. Ritesh Kumar, A.K.O.: KMI-Panlingua at HASOC 2019: SVM vs BERT for Hate Speech and Offensive Content Detection. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
32. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In: Proceedings of the Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC). Bochum, Germany (2016)
33. Saha, B.N., Senapati, A.: CIT Kokrajhar Team: LSTM based Deep RNN Architecture for Hate Speech and Offensive Content (HASOC) Identification in Indo-European Languages . In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
34. Saha, P., Mathew, B., Goyal, P., Mukherjee, A.: HateMonitors at HASOC 2019: Language Agnostic Online Abuse Detection. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
35. Salminen, J., Almerekhi, H., Kamel, A.M., Jung, S.g., Jansen, B.J.: Online hate ratings vary by extremes: A statistical analysis. In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. pp. 213–217. ACM (2019)
36. Schmidt, A., Wiegand, M.: A Survey on Hate Speech Detection Using Natural Language Processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics. pp. 1–10. Valencia, Spain (2017)
37. Sreelakshmi.K, P., K.P, S.: AmritaCEN at HASOC 2019: Hate SpeechDetection in Roman and Devanagiri Scripted Text. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
38. Struß, J.M., Siegel, M., Ruppenhofer, J., Wiegand, M., Klenner, M.: Overview of germeval task 2, 2019 shared task on the identification of offensive language (2019)
39. Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., Daelemans, W.: The automated detection of racist discourse in dutch social media. Computational Linguistics in the Netherlands Journal **6**, 3–20 (2016)
40. Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., Daelemans, W.: A dictionary-based approach to racism detection in dutch social media. arXiv preprint arXiv:1608.08738 (2016)
41. Urmi Saha, A.D., Bhattacharyya, P.: IIT Bombay at HASOC 2019: Supervised Hate Speech and Offensive Content Detection in Indo-European Languages. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)
42. Wagner, K., Bumann, C.: Challenges in annotating a corpus for automatic hate speech detection. In: BOBCATSSS Paris, January (2020)
43. Wiegand, M., Ruppenhofer, J., Kleinbauer, T.: Detection of abusive language: the problem of biased datasets. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 602–608 (2019)
44. Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the germeval 2018 shared task on the identification of offensive language (2018)
45. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the Type and Target of Offensive Posts in Social Media. In: Proceedings of NAACL (2019)

46. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666 (2019)
47. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983 (2019)

A Appendix

A.1 Annotation Guidelines for HASOC 2019

HATE SPEECH Ascribing negative attributes or deficiencies to groups of individuals because they are members of a group (e.g. all poor people are stupid). Hateful comment toward groups because of race, political opinion, sexual orientation, gender, social status, health condition or similar.

OFFENSIVE Degrading, dehumanizing or insulting an individual. Threatening with violent acts.

PROFANITY Unacceptable language in the absence of insults and abuse. This typically concerns the usage of swearwords (Scheiße, Fuck etc.) and cursing (Zur Hölle! Verdammt! etc.).

OTHER Normal content, statements, or anything else. If the utterances are considered to be "normal" and not offending to anyone, they should not be labeled. This could be part of youth language or other language registers.

We expect most posts to be OTHER, some to be HATE and the other two categories to be less frequent.

Dubious cases which are difficult to decide even for humans, should be left out.

B Appendix

B.1 Systems and Approaches at HASOC 2019

The following tables summarize the approaches used by the teams. The last col-umn has an entry when the team compared several approaches and clearly identi-fied a best one. The first table shows the approaches which used technology without Deep Learning or for which a traditional approaches performed best. The second table shows the approaches which used Deep Learning.

Table 16. Participants Team Approaches : based on traditional classifiers

Team name	**Affiliation**	**Text Representation and Classifier**	**Best run (when applicable)**
DEEP [24]	Benha Univ., Egypt & Mangalore Univ.	TF/IDF weighting, SVM, MLP	SVM
IRLAB@IITBHU [2]	IIT Varanasi	SVM, Xboost	SVM
A3-108 [23]	IIIT-Hyderabad	Several Word and character n-grams, Length of tweet, SVM, AdaBoost, Random Forest,LSTM	ML instead of Deep Learning
DA_Master [26]	DAIICT	TFIDF, log. Regression;, CNN	TFIDF, log. Regression
DLRG [28]	Vellore Institute of Technology, Chennai	TF/IDF, Ensemble	Random Forest
UACh-INAOE [8]	Univ. Autonoma Chihuahua & Instit. National de Astrofisica	Word and character n-grams, Word embeddings, Clustering, Flesch Scores, NER count, LR, SVM	Word and character n-gram frequencies, LR
Hate Monitors [34]	IIT Kharagpur	BERT, LASER, Light Gradient Boosting	Light Gradient Boosting
Dracarys [41]	IIT Bombay & Apple	CNN, SVM	SVM

Table 17. participants Team Approaches : based on Deep neural model & Transfer Learning

Team name	**Affiliation**	**Text Representation and Classifier**	**Best run (when applicable)**
AI_ML_NIT [17]	IIT Patna	CNN, fastText	fastText, one-Hot
Amrita [37]	Amrita Vishwa Vidyapeetham	CNN, LSTM fast-Text	
LGI2P [13]	Univ Montpellier	fastText	
CIT Kokrajhar [33]	University of Edmonton & CIT Kokrajhar	LSTM	LSTM
YNU_wb [6]	Yunnan Univ.	ON-LSTM, Attention mechanism, K-folding, Ensemble	ON-LSTM
QutNocturnal [4]	Queensland Univ. of Technology, Brisbane	LSTM, DNN,SVM, kNN. Boosting	CNN,Transfer learning
QMUL-NLP [15]	Queen Mary Univ. London	TFIDF, Word2Vec, LSTM	Word2Vec & LSTM
am905771 [20]	IIT Varanasi	Glove, Bi LSTM,Attention	Glove Twitter
IIITG-ADBU [3]	IIIT Guwahati, & IBM Research India	ELMO, GLOVE, fastText, log. Regression, SVM, BiLSTM	fastText , BiLSTM
Vito [25]	Univ. Politécnica de Valencia	POS tagging,CNN, BiLSTM,	Ensemble
TheNorth [27]	Luleå University of Technology, Sweden	Bi-LSTM	Bi-LSTM
FALSE Positive [16]	IIIT Guwahati	Stacked BiLSTM; CNN	
RALIGRAPH [19]	Univ. of Montral	BERT, Graph CNNPre-trained with external Founta corpus	VCGN-BERT
3Idiots [21]	Univ. of Illinois & IIT Kanpur	BERT, All 3 tasks in one	BERT cased
BRUMS [29]	Univ of Wolverhampton, Rochester Institute of Techn. & Birmingham City Univ	Several deep learning architectures including LSTM, GRU, Attention, 2D Convolution,light pre-processing	BERT
KMI-Panlingua [31]	Bhimrao Ambedkar Univ. & Panlingua & Charles Univ. Prague	BERT, Char and word ngrams + SVM	BERT
LSV-UdS [10]	Saarland University	BERT, SVM,External collections	BERT