TUW-Inf at HASOC 2021 - Offensive text detection on English Twitter with deep learning models and rule-based systems

Kinga Gémes, Ádám Kovács, Markus Reichel, Gábor Recski

kinga.gemes@tuwien.ac.at, adam.kovacs@tuwien.ac.at, mx.markus.rei@gmx.net, gabor.recski@tuwien.ac.at

12 Oct 2021

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2019: (Mandl et al., 2019)

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2019: (Mandl et al., 2019)

- First HASOC, inspired by OffensEval of 2019 (Zampieri et al., 2019) and GermEval of 2018 (Wiegand, Siegel, and Ruppenhofer, 2018)

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2019: (Mandl et al., 2019)

- First HASOC, inspired by OffensEval of 2019 (Zampieri et al., 2019) and GermEval of 2018 (Wiegand, Siegel, and Ruppenhofer, 2018)
- Languages: English, Hindi, German

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2019: (Mandl et al., 2019)

- First HASOC, inspired by OffensEval of 2019 (Zampieri et al., 2019) and GermEval of 2018 (Wiegand, Siegel, and Ruppenhofer, 2018)
- Languages: English, Hindi, German
- Subtasks:

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2019: (Mandl et al., 2019)

- First HASOC, inspired by OffensEval of 2019 (Zampieri et al., 2019) and GermEval of 2018 (Wiegand, Siegel, and Ruppenhofer, 2018)
- Languages: English, Hindi, German
- Subtasks:
  - Subtask A: Binary Classification - `HOF` (Hate and Offensive), or `NOT` (Non Hate-Offensive)

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2019: (Mandl et al., 2019)

- First HASOC, inspired by OffensEval of 2019 (Zampieri et al., 2019) and GermEval of 2018 (Wiegand, Siegel, and Ruppenhofer, 2018)
- Languages: English, Hindi, German
- Subtasks:
  - Subtask A: Binary Classification - `HOF` (Hate and Offensive), or `NOT` (Non Hate-Offensive)
  - Subtask B: Fine-grained Classification - `HATE` (Hate speech), `OFFN` (Offenive), or `PRFN` (Profane)

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2019: (Mandl et al., 2019)

- First HASOC, inspired by OffensEval of 2019 (Zampieri et al., 2019) and GermEval of 2018 (Wiegand, Siegel, and Ruppenhofer, 2018)
- Languages: English, Hindi, German
- Subtasks:
  - Subtask A: Binary Classification - `HOF` (Hate and Offensive), or `NOT` (Non Hate-Offensive)
  - Subtask B: Fine-grained Classification - `HATE` (Hate speech), `OFFN` (Offenive), or `PRFN` (Profane)
  - Subtask C: Binary Classification - `TIN` (Targeted Insult), or `UNT` (Untargeted)

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2020: (Mandl et al., 2020)

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2020: (Mandl et al., 2020)

- Languages: English, Hindi, German

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2020: (Mandl et al., 2020)

- Languages: English, Hindi, German
- Subtasks:

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2020: (Mandl et al., 2020)

- Languages: English, Hindi, German
- Subtasks:
  - Subtask A: Binary Classification - `HOF` (Hate and Offensive), or `NOT` (Non Hate-Offensive)

Brief history of Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC)

2020: (Mandl et al., 2020)

- Languages: English, Hindi, German
- Subtasks:
  - Subtask A: Binary Classification - `HOF` (Hate and Offensive), or `NOT` (Non Hate-Offensive)
  - Subtask B: Fine-grained Classification - `HATE` (Hate speech), `OFFN` (Offenive), or `PRFN` (Profane)

HASOC 2021 (Mandl et al., 2021)

- Languages: English, Hindi, Marathi

HASOC 2021 (Mandl et al., 2021)

- Languages: English, Hindi, Marathi
- Subtasks:

HASOC 2021 (Mandl et al., 2021)

- Languages: English, Hindi, Marathi
- Subtasks:
    - Subtask 1:
        - Subtask A: Binary Classification - `HOF` (Hate and Offensive), or `NOT` (Non Hate-Offensive)

HASOC 2021 (Mandl et al., 2021)

- Languages: English, Hindi, Marathi
- Subtasks:
  - Subtask 1:
    - Subtask A: Binary Classification - `HOF` (Hate and Offensive), or `NOT` (Non Hate-Offensive)
    - Subtask B: Fine-grained Classification - `HATE` (Hate speech), `OFFN` (Offenive), or `PRFN` (Profane)
  - Subtask 2: Binary Classification of comments in a comment-chain in Code-Mixed languages

HASOC 2021 (Mandl et al., 2021)

Participants:

- **English Subtask 1A**: 56 teams
- **English Subtask 1B**: 37 teams
- Hindi Subtask 1A: 34 teams
- Hindi Subtask 1B: 24 teams
- Marathi Subtask 1A: 25 teams
- Subtask 2: 15 teams & 1 baseline system

Our methods

Binary Classification:

- Simple BERT-based (Devlin et al., 2019) method using *bert-base-uncased*, preprocessing, and balanced weighted loss
- Rule-based method using active learning and AMR (Banarescu et al., 2013) graphs
- Union voting: if any one of the above systems says that the text is offensive, it is offensive
- Logreg voting: trained on the probabilistic output of the BERT system and the binary output of the rule-based system

Our methods

Fine-Grained Classification:

- Random Forest Classification using n-gram features (n=3)
- Random Forest Classification using AMR edge features
- BERT based method, trained on each category as binary classification
- Soft voting: summarizing the output probabilities of the two random forest classifiers with similar weights, then taking the argmax of the probabilities

Results on HASOC 2021 English 1A

	HOF			NOT			macro avg		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Top	85	91	**88**	83	73	78	84	**82**	**83**
Logreg voting	81	93	87	85	65	74	83	79	80
BERT	80	95	87	89	61	73	**85**	78	80
Union voting	80	**96**	87	89	60	72	**85**	78	79
Rule	**87**	45	59	50	**89**	64	68	67	62

Results on HASOC 2021 English 1B

	HATE			OFFN			PRFN			NONE		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BERT	50	**62**	**55**	42	**59**	**49**	**70**	73	72	**89**	62	**73**
n-gram	**55**	42	48	44	31	37	**70**	**86**	**77**	68	71	69
soft voting	52	31	39	**53**	26	35	69	**86**	76	62	**75**	68
AMR	20	10	14	26	10	15	53	50	51	43	65	51

	macro avg		
	P	**R**	**F**
Top	**67**	**66**	**67**
BERT	63	64	62
n-gram	59	57	58
soft voting	59	54	55
AMR	35	34	33

Qualitative analysis on the 2019-2020 data

ID	Text
FP1*	Yeah, so, Islam is an idea, not a race. ...a terrible, hateful, idea and you disgrace yourself in its defense. Everybody likes the legal immigrants. Most of them are voting Trump. :) Sincerely, some guy, native of some place.

ID	Text
FP2	Definition of FOOL as per http://dictionary.com , http://en.oxforddictionaries.com “a silly or stupid person; a person who lacks judgment or sense / a clown” Why @nipfp_org_in @bsindia etc are tolerating him?
FP8	The Twitter troll army has gone so mad that they’re uninstalling Zomato to show their support knowing that they’ll install it again after few days when lured by offers and discounts. I want to ask dumb Twitterians isn’t this hypocrisy?

Qualitative analysis on the 2019-2020 data

ID	Text
FP3	@ConnorJaundrell So damn gooood
FP4	If a Hindu mob were to lynch Muslim doctors, every political leader, Modi to Mamata to Rahul to Nitish to Kejri would announce a slew of compensation, visit the victims, and do elaborate Man ki Baats. Mujhe Maro, Mera Doctors ko maat maro, maat maro mere bhaiyon! Hindus disowned
FP5	Mamata Banerjee totally lost Plot in dealing with the Doctors. Immediate action could have prevented crisis. But BJP and Bhakts should refrain from commenting on Mamata's Statements. In similar situation, BJP MOS wanted protesting doctors to become Naxals so that they can be shot
FP6	An image of arrested sex racketeers is going viral with different false claims. One of the false claim says insinuates that 'Hindu terrorists' were caught in Allahabad. Another false claim says that Rohingya Muslims were kidnapping boys in MP... https://timesofindia...

Qualitative analysis on the 2019-2020 data

ID	Cat.	Text
FP3		@ConnorJaundrell So damn gooood
FN10	PRFN	RT @animatrocities: My sister doesn't have a twitter, she's only 14 but dammmmn she's gonna go far!! https://t.co/N16K1VsT7i

Qualitative analysis on the 2019-2020 data

ID	Cat.	Text
FN1	HATE	world Muslim population explosion is real & will permanently change course of human civilization. If BJP not take it biggest challenge n control the population, Dark days r Ahead. What has not happened in last 10K yrs will happen in next 50 years! @AmitShah @narendramodi
FN2	HATE	Yeah especially when there will be 2 Muslim political party in the state against one secular party BJP.
FN3	HATE	That's not white supremacy though is it. That means whites thinking AND ACTING as though they are superior to others. What these Jew/Muslim joint actions are concerned with is protecting stuff like Halal/Kosher, circumcision and the right to wear religious garb. 1/2

Qualitative analysis on the 2019-2020 data

ID	Cat.	Text
FN6	OFFN	RT @JenniThorburn: It's pure privilege to watch #Eurovision and ignore Palestinian people. You're emboldening an apartheid regime. You're...
FN9	OFFN	Best Totichor in the worldproud you sir ❤️❤️❤️

Qualitative analysis on the 2019-2020 data

ID	Cat.	Text
FN4	HATE	@BernieSanders Bernie will never be the President of the United States of America. https://t.co/O0FPo0bTc9
FN8	OFFN	RT @NiteLoops: Kim Kardashian and Kanye West naming their child Psalm West made me think of this video https://t.co/9ouWog1GEO

Qualitative analysis on the 2019-2020 data

ID	Cat.	Text
FN12	HATE	@SkyNews Wherever they go their moronic brain dead fans wreak havoc.
FN16	OFFN	He's always been a fraud, cheat & racist. But this behavior is off the charts insane. #FuckTrump
FN19	PRFN	Wonder if its Mamata/Yesudas tweet or use of 'marketing fuck up' that helpd me lose 15 folowrs yest. Wazzat abt separatng wheat from chaff?

Qualitative analysis on the 2019-2020 data

ID	Cat.	Text
FN13	HATE	#boycottzomato #IStandWithAmit #Zomato you have two rules simultaneously for hindus and for muslims..... And that's not fair.... We live in a democratic secular country where every 1 has the right of choice.... By hurting the majority sentiments, you have taken a risk...Not fair.
FN15	OFFN	U no what, i cant b mad at paul p... i mean the celtics retired this man's number an he won wit them. But ryan smh…https://t.co/LYDCzzytt3
FN17	OFFN	RT @staypufd: He's also one who said they had proof of collusion!!! He's a liar all around, just like Schiff! https://t.co/3znT-bXt1cq

Qualitative analysis on the 2019-2020 data

ID	Cat.	Text
FN14	HATE	Madam ji pls also raise your loud voice against the exodus of Kashmiri pandits by your so called ’ ’ who raped,butchered,acid attacked hindu women just because of their religion..pls also raise your voice against the slogan’kashmir mein rehna hai toh Allah hu akbar
FN18	OFFN	Halala was never part of islam lekin halala ke naam par mulle khoob maje le rabe tumahri maa behan beti ke aur tum bhi maje se dekh rahe unki izzat lutate huye. Aise to baat baat par bomd fod dete ho lekin ek bhi muslim londa inn mullon ke against kuch nahi bolta. Dhikkar hai

Qualitative analysis on the 2019-2020 data

ID	Cat.	Text
FN15	OFFN	U no what, i cant b mad at paul p... i mean the celtics retired this man's number an he won wit them. But ryan smh...https://t.co/LYDCzzytt3
FN20	PRFN	Ass usual seculars... thoughts? Big hearted opinions??? https://t.co/6ViYU01L6K

Bibliography

Banarescu, Laura et al. (2013). "Abstract Meaning Representation for Sembanking". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 178–186. URL: `https://www.aclweb.org/anthology/W13-2322`.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proc. of NAACL*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: `https://aclanthology.org/N19-1423`.

Mandl, Thomas et al. (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 '19. Kolkata, India: Association for Computing Machinery, pp. 14–17. ISBN: 9781450377508. DOI: 10.1145/3368567.3368584. URL: `https://doi.org/10.1145/3368567.3368584`.

Mandl, Thomas et al. (2020). "Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German". In: *Forum for Information Retrieval Evaluation*. FIRE 2020. Hyderabad, India: Association for Computing Machinery, pp. 29–32. ISBN: 9781450389785. DOI: 10.1145/3441501.3441517. URL: `https://doi.org/10.1145/3441501.3441517`.

Mandl, Thomas et al. (2021). "Overview of the HASOC subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages". In: *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*. CEUR. URL: `http://ceur-ws.org/`.

Wiegand, Michael, Melanie Siegel, and Josef Ruppenhofer (2018). "Overview of the GermEval 2018 shared task on the identification of offensive language". In: *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018*. Vienna, Austria: Austrian Academy of Sciences, pp. 1–10. ISBN: 978-3-7001-8435-5. URL: `https://nbn-resolving.org/urn:nbn:de:bsz:mh39-84935`.

Zampieri, Marcos et al. (2019). "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 75–86. DOI: 10.18653/v1/S19-2010. URL: `https://aclanthology.org/S19-2010`.