Overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts

Parth Patwa[1(✉)], Mohit Bhardwaj[2], Vineeth Guptha[4], Gitanjali Kumari[3], Shivam Sharma[2,4], Srinivas PYKL[1], Amitava Das[4], Asif Ekbal[3], Md Shad Akhtar[2], and Tanmoy Chakraborty[2]

[1] IIIT Sri City, Sri City, India
{parthprasad.p17,srinivas.p}@iiits.in
[2] IIIT Delhi, Delhi, India
{mohit19014,shad.akhtar,tanmoy}@iiitd.ac.in,
shivam.sharma23@wipro.com
[3] IIT Patna, Patna, India
{gitanjali_2021cs,asif}@iitp.ac.in
[4] Wipro Research, Bangalore, India
{bodla.guptha,amitava.das2}@wipro.com

Abstract. Fake news, hostility, defamation are some of the biggest problems faced in social media. We present the findings of the shared tasks (https://constraint-shared-task-2021.github.io/) conducted at the CONSTRAINT Workshop at AAAI 2021. The shared tasks are 'COVID19 Fake News Detection in English' and 'Hostile Post Detection in Hindi'. The tasks attracted 166 and 44 team submissions respectively. The most successful models were BERT or its variations.

Keywords: Fake news · COVID-19 · Hostility · Hindi · Machine learning

1 Introduction

A broad spectrum of harmful online content is covered under the umbrella of Hostile communication over social media. Currently, more than $1/3^{rd}$ of the population of the world's two biggest democracies USA [31] and India [37], subscribe to social media-based information. This places these platforms as prime sources of information consumption, in the form of news articles, marketing advertisements, political activities, etc. While the engagement of users on social media was touted as a healthy activity when it started gaining prominence, public behavior now seems to be inducing significant negativity in terms of hostile information exchange primarily in the form of hate-speech, fake-news, defamation, and offense [57]. The problem is magnified by what is termed as the *hostile-media effect* which establishes the perception bias for a common piece of information, that gets induced within the minds of users of one ideological stand-point against that of another [68], effectively pitting social media users constantly at odds.

T. Chakraborty et al. (Eds.): CONSTRAINT 2021, CCIS 1402, pp. 42–53, 2021.
https://doi.org/10.1007/978-3-030-73696-5_5

In particular, dissemination of *spurious* content has been taking its own course of nourishment for quite some time, but the usage of the term fake news is relatively new in this context. It was towards the end of the 19^{th} century that a major daily published "Secretary Brunnell Declares Fake News About His People is Being Telegraphed Over the Country" [3]. Today, this term has become a house-hold entity, be it a daily waged employee or the head of a state [72], usually to bring forth the context of an idea that has in some ways blown out of proportion. Fake news within the context of COVID-19, the outbreak that has led countries scrambling for medical and other resources, has increased the threat significantly. Even global organizations like WHO are not spared of the consequences of such malicious phenomenon [2]. The rampant dissemination of fake news about COVID-19 and other topics on social media not only leads to people being misled but consequently threatens the very fiber of a healthy society and eventually democracies. For the democratic values to be upheld and the power of making the right conclusion to be vested with people, effective mechanisms need to be in place for facilitating scrutinized knowledge [56].

Social media has now become a platform for news-aggregation by presenting content in a source-agnostic manner. This paves way for content delivery, which is politically biased, unreliable, fact-check worthy, and stemming from the ill-intentions of malicious online trolls, cyber-criminals, and propaganda agencies, to influence the reader's perception towards pre-defined ideas, effectively inducing hostility and chaos within a democratically free social environment. This is amplified by the constant exposure to a static ecosystem of digital information, that people tend to believe as true over a period of time [29]. Such situations need thorough fact-verification, that most people ignore [1].

This paper describes the details of shared tasks on *COVID19 Fake News Detection in English* and *Hostile Post Detection in Hindi* which were organized jointly with the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT) at AAAI 2021.

2 Related Work

Fake news is information that is created false intentionally to deceive the readers. It is used to mislead readers and spread misinformation on topics such as politics, religion, marketing, and finance [16]. On the other hand, hostile posts are abusive, hateful, sarcastic, and aggressive content on social media. The diffusion of fake news and hostile information leads the reader astray from facts, which negatively affects the harmony of the society and mental health of social media users [10,17]. Researchers have claimed that the spread of fake and hostile information on social media affects the prestige of many organizations and individuals [15] and gives mental and emotional stress to the victim [10]. Fake news might affect the opinion of the customer by influencing them to buy products from the market based on the fake reviews and news on social media, which can be considered as a type of cybercrime [45]. Hate speech is used as a negative behavior on social media to put mental stress on the victim; this can include

attacks on religious groups, defaming the user, or other types of cyberbullying activities that could be beyond offensive content on social networks [19].

Many researchers are working constantly to build a more robust automated fake content detection system. Workshops and shared tasks like pan2020 [5], Deepfake challenge [22], Fakeddit, [49] etc. were conducted to draw researchers' interest in this area. Few interfaces like [65] which can collect fake news networks for a given post from Twitter were created. Researchers have tried to develop fact-checking algorithms [69] and BERT based models [36] to detect fake news.

There is abundant work is going on in the field of hostile information detection. Many datasets on hostile content are publicly available [41,46,47,59]. Four workshops [4,24,55,71] on abusive language detection were conducted from 2017–2020. The TRAC1 [39] and TRAC2 [40] shared tasks aimed at detecting social media aggression in Hindi and English. Chakravarthi et al. 2021 [13] organized a shared task on offensive language detection in Dravidian languages.

In addition to the above works, researchers have also been trying to discover algorithms to identify hostile content. Among other techniques, Deep learning (CNN, LSTM) [6,50] and BERT based models [54,58] have been quite popular.

3 COVID-19 Fake News Detection in English

The fake news detection shared task is a binary classification problem. The objective is to identify whether a given English post is fake news or real news.

COVID-19 Fake News Dataset: The dataset consists of a total of 10700 English posts out of which 5600 are real news. The Real news is collected from verified Twitter handles and gives useful information regarding COVID19. Fake news consists of claims that are verified to be false. Fake News posts are collected from various social media platforms such as Twitter, Facebook, Whatsapp and from fact-checking websites such as Politifact, NewsChecker, Boomlive, etc. All annotations were done manually. For more details, please refer [51].

Examples of Fake News

- Dr. Fauci: Paint Gums of Covid-19 Carriers Purple And Give Them A Laxative https://t.co/kuCWJyE2Bq #donaldtrump #coronavirus #andywarhol
- Assassination of the Tunisian doctor Mahmoud Bazarti after his announcement of finding a successful vaccine for COVID-19 in Germany.

Examples of Real News

- Growing evidence suggests #COVID19 can spread before people show symptoms (pre-symptomatic) and from people who have #coronavirus but never show symptoms (asymptomatic). Cloth face coverings help prevent spread of COVID-19 in these situations. See Q&A: https://t.co/vuYx19NZPE. https://t.co/RE9K3kZmYR
- Risk of secondary COVID transmission is about 10% at home new contact tracing study finds. https://t.co/olhnVaLf29

Evaluation: The submissions are ranked according to their weighted average F1 score. F1 score is calculated for each class and the average is weighted by the number of true instances for that class. We also calculate the precision, recall, and accuracy. The participants were asked to submit at most 5 runs on the test set and the best run was considered for the leaderboard.

Baseline Models: To give the reference score for the participants we provided baseline models. The preprocessing step involves the removal of links, stopwords, non-alphanumeric characters. TF-IDF scores were used to select features and ML models like logistic regression, support vector machine (SVM), etc. were used. SVM performs the best and achieves an F1-score of 93.32%. For more details please refer to [51].

4 Hostile Post Detection in Hindi

The Hindi hostility detection shared task focuses on detecting the presence of hostile content in Hindi social media posts. There are two sub-tasks - Coarse-grained hostility detection and fine-grained hostility detection. Coarse-grained includes binary classification of a post into Hostile vs Non-Hostile. Fine-grained sub-task includes multi-label classification of hostile posts into one or more of the four hostile dimensions: fake news, hate speech, offensive, and defamation.

Data: The dataset consists of 8192 texts in Hindi from various social media platforms like Twitter, Facebook, WhatsApp, etc. A post can be either non-hostile or can belong to one or more of the four hostile classes - fake, hate, offensive, and defamation. 3834 texts are hostile and the remaining 4358 are non-hostile. Within the fine-grained hostile dimensions, the number of samples for defamation, fake, hate, and offensive are 810, 1638, 1132, and 1071 respectively. For more details please refer [11]. Data collection Summary:

- *Fake News:* Popular fact-checking websites such as BoomLive[1], Dainik Bhaskar[2], etc. were used to collect topics for fake news which were then manually searched overall popular social media platforms and carefully annotated.
- *Hate Speech:* A list of users posting or encouraging tweets that are violent towards minorities based on religion, race, ethnicity, etc. was curated and their timelines were tracked to get more hateful posts. From their timelines, similar users whose hateful content they are sharing were also tracked.
- *Offensive Posts:* Twitter API[3] was used to query a list of most common swear words in Hindi which were curated by [32].
- *Defamation Posts:* Viral news articles regarding defamation of either an individual or an organization are studied to decide the reality of the situation and then posts regarding similar topics were searched on all popular social media platforms and correctly annotated.

[1] https://hindi.boomlive.in/fake-news.
[2] https://www.bhaskar.com/no-fake-news/.
[3] https://developer.twitter.com/en/docs/twitter-api.

– *Non-Hostile Posts:* Majority of the samples are collected through popular trusted sources like BBCHindi. These samples are manually checked to ensure that their content does not belong to any of the four hostile dimensions. Non-verified users also contribute to around 15% of the total non-hostile samples.

Examples:

Defamation, offensive: #JNU में हुई #तोड़फोड़ के बाद #गर्ल्स हॉस्टल में #बिखरा हुआ #सामान #धन्य हैं यहां की स्टूडेंट। बहुत दुख हुआ इन लोगों की बुक्स देखकर सब फट गई है।

Offensive: @User ये स#ला टिकट ब्लैकिया इतनी हिम्मत लाता कहाँ से है

Hate, offensive: RT @User: पिछले 6 वर्षों ने यह सिद्ध कर दिया कि कांग्रेस कोई राजनैतिक दल नही.... एक छुपा हुआ इस्लामिक संगठन है....

Fake: बिहार चुनाव में प्रचार करेंगी कंगना रनौत/स्कूल कॉलेज रहेंगे बंद

Defamation, offensive: User1 User2 अर्पित का छोटा है। वाया – पड़ोस वाली कु#या

Non-hostile: स्पेशल फ़्रंटियर फ़ोर्स के कमांडो नीमा तेंज़िन की अंतिम यात्रा में लगे भारत माता की जय के नारे...

Evaluation: All the submissions are ranked separately for both the sub-tasks. For the coarse-grained sub-task, the weighted average F1 score for hostile and non-hostile classes was used for evaluation. For the fine-grained sub-task, we take the weighted average of F1 scores of each of the four hostile dimensions. The participants were asked to submit at most 5 runs on the test set and the best run was considered for the leaderboard.

Baseline: We use one vs all strategy for multi-label classification. We train 5 models for each label in a binary fashion. For each classifier, m-BERT[4] model is used to extract post embeddings. The last encoder layer of m-BERT gives 768-dimensional word embeddings. The mean of word embeddings for every word in the post is used to represent the entire post embedding. ML-based classifiers are trained on these embeddings. SVM performed better than Logistic Regression, Random Forest, and Multi-Layer Perception. For fine-grained classifiers, only hostile samples are used for training to handle class imbalance. Our baseline achieves a weighted F1-Score of 84.22% for coarse-grained sub-task and a weighted average F1-score of 54.2% for fine-grained sub-task on the test set. For more details, please refer [11].

5 Participation and Top Performing Systems

Total 166 teams participated in the fake news detection task whereas 44 teams participated in the Hindi hostile post detection task. 52 teams submitted a system description paper across both the tasks. 18 papers were accepted for publications and 10 papers were accepted as non-archival papers. All the accepted papers and the corresponding tasks they participated in are provided in Table 1.

[4] https://huggingface.co/bert-base-multilingual-uncased.

5.1 Winning Systems

- **g2tmn**[25] achieved the best results on the fake news detection task. They preprocess the data by removing URLs, converting emojis to text, and lowercasing the text. Their system is an ensemble of 3 CT-BERT models [48].
- **IREL_IIIT** [53] achieved the best results on the coarse-grained sub-task of the Hostility detection task. They use 3 feature pipelines - cleaned text, hashtags, and emojis. IndicBERT [34] trained using Task Adaptive Pretraining (TAPT) [28] approach is used to extract contextual information from the text. Finally, the representations of the 3 pipelines are concatenated and given to a classification layer.
- **Zeus** [73] achieved the best results on the fine-grained sub-task of the hostility detection task. They use ensemble of 5 BERT [21] models.

Table 1. Accepted papers and the corresponding tasks that they participated in. Out of 52, 18 papers were accepted for archival publication and 10 papers were accepted as non-archival. Total 5 papers report results on both the tasks. (English - COVID-19 Fake News Detection in English, Hindi - Hostile Post Detection in Hindi).

Paper	Task	Archival
Ben Chen et al. 2021 [14]	English	Yes
Arkadipta De et al. 2021 [20]	Hindi	Yes
Azhan and Ahmad 2021 [7]	English, Hindi	Yes
Zutshi and Raj 2021 [74]	English	Yes
Xiangyang Li et al. 2021 [42]	English	Yes
Kamal, Kumar and Vaidhya 2021 [35]	Hindi	Yes
Glazkova, Glazkov and Trifinov 2021 [25]	English	Yes
Yejin Bang et al. 2021 [8]	English	Yes
Siva Sai et al. 2021 [60]	Hindi	Yes
Baris and Boukhers [9]	English	Yes
Tathagata Raha et al. 2021a [53]	Hindi	Yes
Varad Bhatnagar et al. 2021 [12]	Hindi	Yes
Liu and Zhou 2021 [43]	English, Hindi	Yes
Koloski, Stepišnik-Perdih and Škrlj 2021 [38]	English	Yes
Apurva Wani et al. 2021 [70]	English	Yes
Das, Basak and Datta 2021 [18]	English	Yes
Venktesh, Gautam and Masud 2021 [67]	English	Yes
Zhou, Fu and Li 2021 [73]	English, Hindi	Yes
Sharif, Hossain and Hoque 2021 [62]	English, Hindi	No
Gundapu and Mamidi 2021 [26]	English	No
Ramchandra Joshi et al. 2021 [33]	Hindi	No
Thomas Felber 2021 [23]	English	No
Chander Shekar et al. 2021 [63]	Hindi	No
Shifath, Khan and Islam [64]	English	No
Tahtagata Raha et al. 2021b [52]	English	No
Shushkevich and Cardiff 2021 [66]	English	No
Sarthak et al. 2021 [61]	Hindi	No
Ayush Gupta et al. 2021 [27]	English, Hindi	No

5.2 Interesting Systems

Ben Chen et al. 2021 [14] use an ensemble of RoBERTa [44] and CT-BERT [48]. They use heated softmax loss and adversarial training to train their system.

Azhan and Ahmad 2021 [7] a propose layer differentiated training procedure for training ULMFiT [30] model to identify fake news and hostile posts.

Baris and Boukhers 2021 [9] propose ECOL framework that encodes content, prior knowledge, and credibility of sources from the URL links in the posts for the early detection of fake news on social media.

Das, Basak, and Dutta 2021 [18] use a soft voting ensemble of multiple BERT-like models. They augment their system with heuristics which take into account usernames, URLs, and other corpus features along with network-level features.

6 Results

Table 2. Top 15 systems for the English Fake-News Shared task. The systems are ranked by the Weighted F1 score. We report Accuracy, Precision, Recall (R), and weighted F1 score.

Rank	System	Accuracy	Precision	Recall	F1-Score
1	g2tmn	**98.69**	**98.69**	**98.69**	**98.69**
2	saradhix	98.64	98.65	98.64	98.65
3	xiangyangli	98.6	98.6	98.6	98.6
4	Ferryman	98.55	98.56	98.55	98.55
5	gundapusunil	98.55	98.55	98.55	98.55
6	DarrenPeng	98.46	98.47	98.46	98.46
7	maxaforest	98.46	98.47	98.46	98.46
8	dyh930610	98.36	98.37	98.36	98.36
9	abhishek17276	98.32	98.34	98.32	98.32
10	souryadipta	98.32	98.34	98.32	98.32
11	cean	98.27	98.27	98.27	98.27
12	LucasHub	98.32	98.34	98.32	98.32
13	isha	98.32	98.34	98.32	98.32
14	ibaris	98.32	98.34	98.32	98.32
15	Maoqin	98.32	98.34	98.32	98.32
115	**Baseline**	93.32	93.33	93.32	93.42

Table 2 shows the results of the top 15[5] systems for the fake news detection task. All of them are very close to each other and lie between 98.3% and 98.7% F1

[5] Results for all the teams is available at https://competitions.codalab.org/competitions/26655#learn_the_details-result.

score. The winners achieve 98.69% F1 score. For all the systems, there is very little difference between precision and recall. Out of 166 teams, 114 teams were able to beat the baseline whereas 52 could not.

Table 3. Top 10 coarse-grained (CG) systems for the Hindi Hostile posts task. Each system also has a rank for the fine-grained (FG) sub-task. We also report the F1 score for each Fine-grained class.

CG Rank	System	CG F1	Defamation F1	Fake F1	Hate F1	Offensive F1	FG F1	FG Rank
1	IREL_IIIT	**97.16**	44.65	77.18	**59.78**	58.80	62.96	3
2	Albatross	97.10	42.80	81.40	49.69	56.49	61.11	9
3	Quark	96.91	30.61	79.15	42.82	56.99	56.60	19
4	Fantastic_Four	96.67	43.29	78.64	56.64	57.04	62.06	6
5	Aaj Ki Nakli Khabar	96.67	42.23	77.26	56.84	59.11	61.91	7
6	Cean	96.67	44.50	78.33	57.06	**62.08**	63.42	2
7	bestfit_ai	96.61	31.54	**82.44**	58.56	58.95	62.21	5
8	Zeus	96.07	**45.52**	81.22	59.10	58.97	**64.40**	**1**
9	Monolith	95.83	42.0	77.41	57.25	61.20	62.50	4
10	Team_XYZ	95.77	35.94	74.41	50.47	58.29	58.06	16
32	**Baseline**	84.22	39.92	68.69	49.26	41.98	54.20	23

A total of 44 teams participated in the Hindi Hostility Detection Shared task. These are evaluated for both sub-tasks separately. Table 3 shows the results of top the 10[6] systems for the hostility detection task.

- *Coarse-Grained Results:* 31 teams out of 44 surpassed the baseline score of 84.22% weighted F1-score. The submissions range from 97.15% and 29.0% weighted F1-score for this sub-task, with 83.77% and 87.05% weighted F1-Score for the mean and median.
- *Fine-Grained Results:* The Fine-grained sub-task was much more difficult than the coarse-grained sub-task as the winners achieve only 64.39% weighted F1-score. 22 teams out of 44 manage to beat the baseline score of 54.2% which is also the median for fine-grained sub-task. The submissions range from 64.39% to 11.77% with an average of 50.12%. 8 out of the top 10 teams for coarse-grained sub-task also manages to be within the top 10 teams for fine-grained sub-task. The mean F1-scores for each hostile dimension i.e. fake news, hate, offensive, and defamation are 63.05%, 43.74%, 51.51%, and 31.59% respectively. Fake news is the easiest dimension to detect. The defamation class accounts for the lowest average F1 scores due to the lowest number of samples for training.

[6] Results for all the teams is available at https://competitions.codalab.org/competitions/26654#learn_the_details-submission-details.

7 Conclusion and Future Work

In this paper, we describe and summarize the 'COVID-19 Fake News Detection in English' and the 'Hostile Post Detection in Hindi' shared tasks. We see that domain-specific fine-tuning of pre-trained BERT-based models are very successful in both the tasks and is used by the winners and many participants. Ensemble techniques are also quite successful. We saw some interesting methods which are worth exploring further. From the results of fine-grained hostility detection, we can conclude that it is a difficult task and the systems need further analysis and improvement. The shared tasks reported in this paper aim to detect fake news and hostile posts, however, these problems are far from solved and require further research attention.

Future work could involve creating datasets for more languages and providing an explanation of why the post is fake/hostile. Another direction could be to provide the levels of hostility instead of simple yes/no.

References

1. A brief history of fake news. https://www.cits.ucsb.edu/fake-news/brief-history
2. Fake news alert. https://www.who.int/india/emergencies/coronavirus-disease-(covid-19)/fake-news-alert
3. How is 'fake news' defined, and when will it be added to the dictionary?. https://www.merriam-webster.com/words-at-play/the-real-story-of-fake-news
4. Akiwowo, S., et al. (eds.): Proceedings of the Fourth Workshop on Online Abuse and Harms. Association for Computational Linguistics (2020)
5. Arampatzis, A., et al. (eds.): 11th International Conference of the CLEF Association (CLEF 2020). LNCS (2020)
6. Aroyehun, S.T., Gelbukh, A.: 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) (2018)
7. Azhan, M., Ahmad, M.: LaDiff ULMFiT: a layer differentiated training approach for ULMFiT. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 54–61, Springer, Cham (2021)
8. Bang, Y., et al.: Model generalization on COVID-19 fake news detection. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 128–140, Springer, Cham (2021)
9. Baris, I., Boukhers, Z.: ECOL: early detection of COVID lies using content, prior knowledge and source information. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 141–152, Springer, Cham (2021)
10. Beran, T., Li, Q.: Cyber-harassment: a study of a new method for an old behavior. JECR **32**(3), 265 (2005)
11. Bhardwaj, M., et al.: Hostility detection dataset in Hindi (2020)
12. Bhatnagar, V., et al.: Divide and conquer: an ensemble approach for hostile post detection in Hindi. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 244–255, Springer, Cham (2021)
13. Chakravarthi, B.R., et al.: Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages (2021)

14. Chen, B., et al.: Transformer-based language model fine-tuning methods for COVID-19 fake news detection. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 83–92, Springer, Cham (2021)
15. Cheng, Y., Chen, Z.F.: The influence of presumed fake news influence: examining public support for corporate corrective response, media literacy interventions, and governmental regulation. Mass Commun. Soc. **23**(5), 705–729 (2020)
16. Claire Wardle, H.D.: Information disorder: toward an interdisciplinary framework for research and policy making (2017). https://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf
17. Cui, L., Lee, D.: CoAID: COVID-19 healthcare misinformation dataset (2020)
18. Das, S.D., Basak, A., Dutta, S.: A heuristic-driven ensemble framework for COVID-19 fake news detection. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 164–176, Springer, Cham (2021)
19. Davidson, T., et al.: Automated hate speech detection and the problem of offensive language. In: Proceedings of ICWSM (2017)
20. De, A., et al.: Coarse and fine-grained hostility detection in Hindi posts using fine tuned multilingual embeddings. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 201–212, Springer, Cham (2021)
21. Devlin, J., et al.: BERT: pre-training of deep bidirectional transformers for language understanding (2019)
22. Dolhansky, B., et al.: The deepfake detection challenge (DFDC) dataset (2020)
23. Felber, T.: Constraint 2021: machine learning models for COVID-19 fake news detection shared task (2021)
24. Fišer, D., et al. (eds.): Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) (2018)
25. Glazkova, A., Glazkov, M., Trifonov, T.: g2tmn at constraint@AAAI2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 116–127, Springer, Cham (2021)
26. Gundapu, S., Mamidi, R.: Transformer based automatic COVID-19 fake news detection system (2021)
27. Gupta, A., et al.: Hostility detection and COVID-19 fake news detection in social media (2021)
28. Gururangan, S., et al.: Don't stop pretraining: adapt language models to domains and tasks (2020)
29. Holone, H.: The filter bubble and its effect on online personal health information. Croatian Med. J. **57**, 298 (2016)
30. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification (2018)
31. Humprecht, E., Hellmueller, L., Lischka, J.A.: Hostile emotions in news comments: a cross-national analysis of Facebook discussions. Soc. Media+ Soc. **6**(1), 2056305120912481 (2020)
32. Jha, V.K., et al.: DHOT-repository and classification of offensive tweets in the Hindi language. Procedia Comput. Sci. **171**, 2324–2333 (2020)
33. Joshi, R., Karnavat, R., Jirapure, K., Joshi, R.: Evaluation of deep learning models for hostility detection in Hindi text (2021)
34. Kakwani, D., et al.: IndicNLPSuite: monolingual corpora. In: Findings of EMNLP, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages (2020)

35. Kamal, O., Kumar, A., Vaidhya, T.: Hostility detection in Hindi leveraging pre-trained language models. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 213–223, Springer, Cham (2021)
36. Kar, D., Bhardwaj, M., Samanta, S., Azad, A.P.: No rumours please! A multi-indic-lingual approach for COVID fake-tweet detection. arXiv:2010.06906 (2020)
37. Keelery, S.: Social media users in India, October 2020. https://www.statista.com/statistics/278407/number-of-social-network-users-in-india/
38. Koloski, B., Stepišnik-Perdih, T., Škrlj, B.: Identification of COVID-19 related fake news via neural stacking. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 177–188, Springer, Cham (2021)
39. Kumar, R., et al.: Benchmarking aggression identification in social media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018) (2018)
40. Kumar, R., et al.: Evaluating aggression identification in social media. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (2020)
41. Leite, J.A., et al.: Toxic language detection in social media for Brazilian Portuguese: new dataset and multilingual analysis (2020)
42. Li, X., et al.: Exploring text-transformers in AAAI 2021 shared task: COVID-19 fake news detection in English. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 106–115, Springer, Cham (2021)
43. Liu, R., Zhou, X.: Extracting latent information from datasets in the constraint-2020 shared task on the hostile post detection. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 62–73, Springer, Cham (2021)
44. Liu, Y., et al.: RoBERTa: arobustly optimized BERT pretraining approach (2019)
45. Martens, D., Maalej, W.: Towards understanding and detecting fake reviews in app stores. Empirical Softw. Eng. **24**(6), 3316–3355 (2019)
46. Mathew, B., et al.: HateXplain: a benchmark dataset for explainable hate speech detection (2020)
47. Mollas, I., et al.: Ethos: an online hate speech detection dataset (2020)
48. Müller, M., Salathé, M., Kummervold, P.E.: COVID-Twitter-BERT: a natural language processing model to analyse COVID-19 content on Twitter (2020)
49. Nakamura, K., Levy, S., Wang, W.Y.: r/Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection (2020)
50. Nikhil, N., et al.: LSTMs with attention for aggression detection. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (2018)
51. Patwa, P., et al.: Fighting an infodemic: COVID-19 fake news dataset. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 21–29, Springer, Cham (2021)
52. Raha, T., et al.: Identifying COVID-19 fake news in social media (2021)
53. Raha, T., et al.: Task adaptive pretraining of transformers for hostility detection. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 236–243, Springer, Cham (2021)
54. Risch, J., Krestel, R.: Bagging BERT models for robust aggression identification. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (2020)
55. Roberts, S.T., et al. (eds.): Proceedings of the Third Workshop on Abusive Language Online. Association for Computational Linguistics (2019)

56. Rose, J.: To believe or not to believe: an epistemic exploration of fake news, truth, and the limits of knowing. Postdigital Sci. Educ. **2**, 202–216 (2020)
57. Rowe, I.: Deliberation 2.0: comparing the deliberative quality of online news user comments across platforms. J. Broadcast. Electron. Media **59**(4), 539–555 (2015)
58. Safi Samghabadi, N., Patwa, P., PYKL, S., Mukherjee, P., Das, A., Solorio, T.: Aggression and misogyny detection using BERT: a multi-task approach. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (2020)
59. Saha, P., Mathew, B., Goyal, P., Mukherjee, A.: Hateminers: detecting hate speech against women (2018)
60. Sai, S., et al.: Stacked embeddings and multiple fine-tuned XLM-roBERTa models for enhanced hostility identification. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 224–235, Springer, Cham (2021)
61. Sarthak, Shukla, S., Mittal, G., Arya, K.V.: Detecting hostile posts using relational graph convolutional network (2021)
62. Sharif, O., Hossain, E., Hoque, M.M.: Combating hostility: COVID-19 fake news and hostile post detection in social media (2021)
63. Shekhar, C., et al.: Walk in wild: an ensemble approach for hostility detection in Hindi posts (2021)
64. Shifath, S.M.S.U.R., Khan, M.F., Islam, M.S.: A transformer based approach for fighting COVID-19 fake news (2021)
65. Shu, K., et al.: Fakenewsnet: a data repository with news content, social context and spatialtemporal information for studying fake news on social media (2019)
66. Shushkevich, E., Cardiff, J.: TUDublin team at constraint@AAAI2021 - COVID19 fake news detection (2021)
67. Gautam, A., Masud, S.: Fake news detection system using XLNet model with topic distributions: constraint@AAAI2021 shared task. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 189–200, Springer, Cham (2021)
68. Vallone, R., Ross, L., Lepper, M.: The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the Beirut massacre. J. Pers. Soc. Psychol. **49**(3), 577–85 (1985)
69. Vijjali, R., Potluri, P., Kumar, S., Teki, S.: Two stage transformer model for COVID-19 fake news detection and fact checking (2020)
70. Wani, A., et al.: Evaluating deep learning approaches for COVID19 fake news detection. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 153–163, Springer, Cham (2021)
71. Waseem, Z., Chung, W.H.K., Hovy, D., Tetreault, J. (eds.): Proceedings of the First Workshop on Abusive Language Online. Association for Computational Linguistics (2017)
72. Wendling, M.: The (almost) complete history of 'fake news', January 2018. https://www.bbc.com/news/blogs-trending-42724320
73. Zhou, S., Fu, R., Li., J.: Fake news and hostile post detection using an ensemble learning model. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 74–82, Springer, Cham (2021)
74. Zutshi, A., Raj, A.: Tackling the infodemic : analysis using transformer based model. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021, CCIS 1402, pp. 93–105, Springer, Cham (2021)