Hostility Detection in Hindi Leveraging Pre-trained Language Models

Ojasv Kamal(✉), Adarsh Kumar, and Tejas Vaidhya

Indian Institute of Technology, Kharagpur, Kharagpur, West Bengal, India

Abstract. Hostile content on social platforms is ever increasing. This has led to the need for proper detection of hostile posts so that appropriate action can be taken to tackle them. Though a lot of work has been done recently in the English Language to solve the problem of hostile content online, similar works in Indian Languages are quite hard to find. This paper presents a transfer learning based approach to classify social media (i.e. Twitter, Facebook, etc.) posts in Hindi Devanagari script as Hostile or Non-Hostile. Hostile posts are further analyzed to determine if they are Hateful, Fake, Defamation, and Offensive. This paper harnesses attention based pre-trained models fine-tuned on Hindi data with Hostile-Non hostile task as Auxiliary and fusing its features for further sub-tasks classification. Through this approach, we establish a robust and consistent model without any ensembling or complex pre-processing. We have presented the results from our approach in CONSTRAINT-2021 Shared Task [21] on hostile post detection where our model performs extremely well with **3rd runner up** in terms of Weighted Fine-Grained F1 Score (Refer Sect. 4.3 for description of Weighted Fine-grained f1-score).

Keywords: Hostility detection · Pre-trained models · Natural language processing · Social media · Hindi language

1 Introduction

Social media is undoubtedly one of the greatest innovations of all time. From connecting with people across the globe to sharing of information and knowledge in a minuscule of a second, social media platforms have tremendously changed the way of our lives. This is accompanied by an ever-increasing usage of social media, cheaper smartphones, and the ease of internet access, which have further paved the way for the massive growth of social media. To put this into numbers, as per a recent report[1], more than 4 billion people around the world now use

[1] https://datareportal.com/reports/digital-2020-october-global-statshot.

Shared Task in CONSTRAINT 2021.

O. Kamal and A. Kumar—Equal Contribution.

T. Chakraborty et al. (Eds.): CONSTRAINT 2021, CCIS 1402, pp. 213–223, 2021.
https://doi.org/10.1007/978-3-030-73696-5_20

social media each month, and an average of nearly 2 million new users are joining them every day.

While social media platforms have allowed us to connect with others and strengthen relationships in ways that were not possible before, sadly, they have also become the default forums for holding high-stakes conversations, blasting polarizing opinions, and making statements with little regard for those within the screenshot. The recent increase in online toxicity instances has given rise to the dire need for adequate and appropriate guidelines to prevent and curb such activities. The foremost task in neutralising them is hostile post detection. So far, many works have been carried out to address the issue in English [18,28] and several other languages [2,16]. Although Hindi is the third largest language in terms of speakers and has a significant presence on social media platforms, considerable research on hate speech or fake content is still quite hard to find. A survey of the literature suggests a few works related to hostile post detection in Hindi, such as [9,25]; however, these works are either limited by inadequate number of samples, or restricted to a specific hostility domain.

A comprehensive approach for hostile language detection on hostile posts, written in Devanagari script, is presented in [1], where the authors have emphasized multi-dimensional hostility detection and have released the dataset as a shared task in Constraint-2021 Workshop. This paper presents a transfer learning based approach to detect Hostile content in Hindi leveraging Pre-trained models, with our experiments based on this dataset. The experiments are subdivided into two tasks, **Coarse Grained task**: Hostile vs. Non-Hostile Classification and **Fine Grained subtasks**: Sub-categorization of Hostile posts into fake, hate, defamation, and offensive.
Our contribution comprises of improvements upon the baseline in the following ways:

1. We fine-tuned transformer based pre-trained, Hindi Language Models for domain-specific contextual embeddings, which are further used in Classification Tasks.
2. We incorporate the fine-tuned hostile vs. non-hostile detection model as an auxiliary model, and fuse it with the features of specific subcategory models (pre-trained models) of hostility category, with further fine-tuning.

Apart from this, we have also presented a comparative analysis of various approaches we have experimented on, using the dataset. The code and trained models are available at this https url[2].

2 Related Work

In this section, we discuss some relevant work in NLP for Pre-Trained Model based Text Classification and Hostile Post Detection, particularly in the Indian Languages.

[2] https://github.com/kamalojasv181/Hostility-Detection-in-Hindi-Posts.git.

Pretrained-Language Models in Text Classification

Pre-trained transformers serve as general language understanding models that can be used in a wide variety of downstream NLP tasks. Several transformer-based language models such as GPT [23], BERT [5], RoBERTa [14], etc. have been proposed. Pre-trained contextualized vector representations of words, learned from vast amounts of text data have shown promising results in the task of text classification. Transfer learning from these models has proven to be particularly useful in tasks where there is a lack of undisputed labeled data and the inability of surface features to capture the subtle semantics in the text as in the case of hate speech [15]. However, all these pre-trained models require large amounts of monolingual corpus to train on. Nonetheless, Indic-NLP [11] and Indic-Transformers [8] have curated datasets, trained embeddings, and created benchmarks for classification in multiple Indian languages including hindi. [10] presented a comparative study of various classification techniques for Hindi, where they have demonstrated the effectiveness of Pre-trained sentence embedding in classification tasks.

Hostile Post Detection

Researchers have been studying hate speech on social media platforms such as Twitter [29], Reddit [17], and YouTube [19] in the past few years. Furthermore, researchers have recently focused on the bias derived from the hate speech training datasets [3]. Among other notable works on hostility detection, Davidson et al. [4] studied the hate speech detection for English. They argued that some words might reflect hate in one region; however, the same word can be used as a frequent slang term. For example, in English, the term 'dog' does not reveal any hate or offense, but in Hindi (ku##a) is commonly referred to as a derogatory term in Hindi. Considering the severity of the problem, some efforts have been made in Non-English languages as well [2,7,16,25]. Bhardwaj et al. [1] proposed a multi-dimensional hostility detection dataset in Hindi which we have focused on, in our experiments. Apart from this, there are also a few attempts at Hindi-English code-mixed hate speech [26].

3 Methodology

In the following subsections, we briefly discuss the various methodologies used in our experiments. Each subsection describes an independent approach used for classification and sub-classification tasks. Our final approach is discussed in Sect. 3.4.

3.1 Single Model Multi-label Classification

In this approach, we treat the problem as a Multi-label classification task. We use a single model with shared parameters for all classes to capture correlations amongst them. We fine tuned the pre-trained BERT transformer model to

get contextualized embedding or representation by using attention mechanism. We experimented with three different versions of pre-trained BERT transformer blocks, namely Hindi BERT (a compressed form of BERT) [6], Indic BERT(based on the ALBERT architecture) [11], and a HindiBERTa model [24]. The loss function used in this approach can be formulated mathematically as:

$$L(\hat{y}, y) = -\sum_{j=1}^{c} y_j log\hat{y_j} + (1 - y_j)log(1 - \hat{y_j})$$

$$J(W^{(1)}, b^{(1)}, ...) = 1/m \sum_{i=1}^{m} L(\hat{y^i}, y^{(i)})$$

where, c is total number of training examples and m is number of different classes (i.e. non-hostile, fake, hate, defamation, offensive).

3.2 Multi-task Classification

In this approach, we considered the classification tasks as a Multi-task Classification problem. As described in Fig. 1(a), we use a shared BERT model and individual classifier layers, trained jointly with heuristic loss. This is done so as to capture correlations between tasks and subtasks in terms of contextualized embeddings from shared BERT model while maintaining independence in classification tasks. We experimented with Indic-BERT and HindiBERTa (we dropped the Hindi BERT model in this approach as the performance was poor compared to the other two models because of shallow architecture). The heuristic loss can be formulated mathematically as:

$$L = l(x, y) = \{l_1, ..., l_N\}^T$$

where,

$$l_n = -w_n[y_n \cdot log\sigma(x_n) + (1 - y_n) \cdot log(1 - \sigma(x_n))]$$

$$L_{total} = L_{(hostile/non-hostile)} + \lambda \cdot 1/N\{L_{(hurt,defame,fake,offensive)}\}$$

if post is Hostile $\lambda = 0.5$ (contributing to fine grain task), otherwise $\lambda = 0$

3.3 Binary Classification

Unlike the previous two approaches, here we consider each classification task as an individual binary classification problem based on fine tuned contextualised embedding. We fine tuned the BERT transformer block and the classifier layer above it using the binary target labels for individual classes. Same as in Multi-task approach, we experimented this approach with Indic-BERT and HindiBERTa. Binary cross-entropy loss used in this approach can be mathematically formulated as follows:

$$L_i(\hat{y}, y) = -\sum_{j=1}^{c} y_j log\hat{y_j} + (1 - y_j)log(1 - \hat{y_j})$$

where, c is total number of training examples and i is number of independent models for each task

3.4 Auxiliary Task Based Binary Sub-classification

Similar to the previous approach, each classification task is considered as an individual binary classification problem. However, as an improvement over the previous approach, we treat the coarse-grained task as an Auxiliary task and then fuse its logits to each of the fine-grained subtasks. The motivation is that a hostile sub-class specific information shall be present in a post only if the post belongs to hostile class [12]. So, treating it as an Auxiliary task allow us to exploit additional hostile class-specific information from the logits of Auxiliary model. The loss function used in this case was same as described in Binary Classification. The model is described in Fig. 1(b).

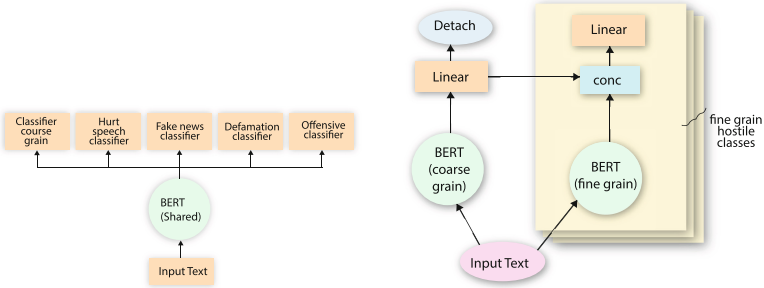

Fig. 1. (a) Multi-task classification model (b) Auxiliary task based binary sub classification model.

4 Experiment

In this section, we first introduce the dataset used and then provide implementation details of our experiments in their respective subsections.

4.1 Dataset Description

As already mentioned in Sect. 1, we evaluate our approach based on the dataset proposed in [1]. As described in the dataset paper, the objective of the task is a classification of posts as Hostile and Non-Hostile and further Multi-label

classification of Hostile posts into *fake*, *hate*, *offensive*, and *defame* classes. The dataset consists of 8192 online posts out of which 4358 samples belong to the non-hostile category, while the rest 3834 posts convey one or more hostile dimensions. There are 1638, 1132, 1071, and 810 posts for fake, hate, offensive, and defame classes in the annotated dataset, respectively. Same as in the paper [1], we split the dataset into 70:10:20 for train, validation, and test, by ensuring the uniform label distribution among the three sets, respectively.

4.2 Pre-processing

Prior to training models, we perform the following pre-processing steps:

- We remove all non-alphanumeric characters except full stop punctuation marks (|, ?) in Hindi, but we keep all stop words because our model trains the sequence of words in a text directly.
- We replace all user mentions and hashtags with a blank space.
- We skip emojis, emoticons, flags etc. from the posts.
- We replace the URLs with the string 'http'.

4.3 Experimental Setup

All the experiments were performed using Pytorch [20] and HuggingFace [30] Transformers library. As the implementation environment, we used Google Colaboratory tool which is a free research tool with a Tesla K80 GPU and 12 GB RAM. Optimization was done using Adam [13] with a learning rate of $1e{-}5$. As discussed earlier in Sect. 3, in our experiments, we used pre-trained HindiBert [6], IndicBert [11] and HindiBERTa [24] Models available in HuggingFace library. Input sentences were tokenized using respective tokenizers for each model, with maximum sequence length restricted to 200 tokens. We trained each classifier model with a batch size of 16. In all the approaches, we used only the first token output provided by each model as input to classifier layer. Each classifier layer has 1 dropout layer with dropout of 0.3 and 1 fully connected layer. Each sub-classification task (fine grained task) was trained only on the hostile labeled examples, i.e. the posts that had at least one label of hostile class, so as to avoid extreme class-imbalance caused by including non-hostile examples. For the evaluation, we have used weighted f1 score [22] as a metric for measuring the performance in both the classification tasks. As suggested in the CONSTRAINT-2021 shared task [21], to measure the combined performance of 4 individual fine-grained sub-tasks together, we have used weighted fine-grained f1 score as the metric, where the weights for the scores of individual classes are the fraction of their positive examples.

Table 1. Results obtained using various methods and models used. Here, **Baseline**: as described in the dataset paper [1], **MLC**: Multi Label Classification, **MTL**: Multitask Learning, **BC**: Binary Classification and **AUX**: Auxiliary Model

Method	Model	Hostile	Defamation	Fake	Hate	Offensive	Weighted
Baseline	–	0.8422	0.3992	0.6869	0.4926	0.4198	0.542
MLC	Hindi-BERT	0.952	0.0	0.7528	0.4206	0.5274	0.4912
	Indic-BERT	0.9581	0.3787	0.7228	0.3094	0.5152	0.513
	HindiBERTa	0.9507	0.3239	0.7317	0.4120	0.4106	0.5122
MTL	Indic-BERT	0.9284	0.0513	0.3296	0.0	0.0	0.1260
	HindiBERTa	0.9421	0.31	0.6647	0.2353	0.5545	0.4738
BC	Hindi-BERT	0.9359	0.130	0.7164	0.47698	0.5388	0.5169
	Indic-BERT	0.9520	0.3030	0.757	0.4745	0.5446	0.5618
	HindiBERTa	0.9421	0.2707	0.6596	0.3175	0.6098	0.4960
AUX	**Indic-BERT**	**0.9583**	**0.42**	**0.7741**	**0.5725**	**0.6120**	**0.6250**
	HindiBERTa	0.9486	0.3855	0.7612	0.5663	0.5933	0.6086

5 Results

In this section, we discuss the results from the different approaches proposed in Sect. 3. Table 1 summarizes the obtained results for different approaches, along with the baseline [1]. Since hostile/non-hostile posts are real phenomenon, we did not perform oversampling and undersampling techniques to adjust class distribution and tried to supply the dataset as realistic as possible. This was done to avoid overfitting (in case of oversampling) and the loss of crucial data (in case of undersampling). As it's clear from Table 1, our best model based on approach described in Sect. 3.4 with Indic-BERT model outperforms the baseline as well as other approaches in both the tasks, i.e. Coarse Grained Task of Hostile vs. Non-Hostile Classification and Fine Grained Task of Hostile Sub-Classification. Moreover, our best model stands as the **3rd** runner up in terms of Weighted fine grained f1 score in the CONSTRAINT-2021 shared task on Hostile Post detection (Results can be viewed here[3]).

6 Error Analysis

Although we have received some interesting results, there are certain dimensions where our approach does not perform as expected. Through this section we try to better understand the obtained f1 scores through some general observations and some specific examples (refer Table 2). Our model did perform comparatively better in fake dimension which implies the model was able to capture patterns in fake samples from dataset to a large extent. However, as can be seen in the example 1, the fake/non-fake classification of posts in certain cases largely

[3] Our team name is **Monolith**.

context/knowledge based. Therefore, in absence of any external knowledge, the method is quite inefficient, particularly in those kind of samples which are under-represented in the dataset. Apart from this, we observe that the defamation scores are the lowest in general. This could be mainly attributed to the overall under-representation of the class in the dataset. Hence a more balanced dataset is critical to boost the defamation f1 score.

Another important observation to note is the existence of metaphorical data in the dataset, which implies meaning different from what semantic information is absent. For example, consider example 2 in the Table 2. This tweet has been inspired by the Hindi idiom "नौ सौ चूहे खा के बिल्ली हज को चली" which means a person after committing every sin in the rule book looks to God for atonement and is used to refer to a hypocritical person indirectly. Such examples lead to mis-classification by models which are primarily based on contextualized embeddings training on simple datasets, as in our case. However, this could be eliminated if the models are pre-trained/fine-tuned on datasets which contain more such examples of metaphorical samples. From our manual inspection, we also observed that the dataset includes some examples, the labels of which are not even apparent to us. For instance, consider example 4. This example simply urges people to speak up and for some cause. Such type of sentence are quite often noticeable in hindi literature. It is impossible to conclude that it is an offensive post with the given data. However, the fact that it is correctly classified by our model reflects bias in the dataset with respect to certain kind of examples, against a generalization of the "Offensive" dimension. Apart from this, we also found some examples which, in our opinion are labeled incorrectly or are possibly ambiguous to be categorised in dimensions being considered. Example 5 means we do not want a favour we only ask for what we deserve which is labeled as defamation however according to us, it is ambiguous to classify it into any

Table 2. Misclassified samples from the dataset

	Post	Annotated Label	Predicted Label
1	हमारे हिन्दू जाट भाईओ पर बोला गहलोत देख लो और वोट दो जाट भाईओ ये साले किसी के सगे नही है	Fake	Not Fake
2	@KanganaTeam नोसो चूहे खाकर कुतिया हज को चली	Defamation	Not Defamation
3	वी डोंट सपोर्ट NRC, CAB, CAA. वापिस जाओ मोदी. टकला अमित गो बैक	Fake	Defamation
4	आज जो आवाज़ नही उठाते वो कल पछतायेंगे,क्यूंकि आज हमारा खामोश रहना ही आने वाली पीढ़ी की गुलामी की ज़मानत है,और लोग आने वाले वक्त में कहेंगे की ज़माना ही खराब था क्या करते,लेकिन नस्लें पूछेंगी तुम खामोश क्यूं थे?	Offensive	Offensive
5	हम किसी से किसी की जागीर नही माँगते हम बस अपने योग्यता के अनुसार अपना हक मांगते हैं	Defamation	Offensive
6	सुनने में आ रहा है कि "ठोको ताली" दोगला दोबारा "बीजेपी" का दरवाजा खटखटा रहा है बीजेपी वालों लात मारो इस कुत्ते को	Not Hate	Hate

of the considered dimensions and largely dependent on the context. Similarly in example 6, someone is being referred as "कुत्ते" which means a dog, according to us it should be hate but is not labeled as hate.

7 Conclusion and Future Work

In this paper, we have presented a transfer learning based approach leveraging the pre-trained language models, for Multi-dimensional Hostile post detection. As the evaluation results indicate, our final approach outperforms baseline, by a significant margin in all dimensions. Furthermore, examining the results shows the ability of our model to detect some biases and ambiguities in the process of collecting or annotating dataset.

There is a lot of scope of improvement for fine Grained with few positive labels. Pre-training on relevant data (such as offensive or hate speech) is a promising direction. In case of Fake news detection, it is very difficult to verify the claim without the use of external knowledge. In future, we would like to extend the approach purposed in paper [27], by using processed-wikipedia knowledge it is possible to significantly improve fake news detection accuracy.

Acknowledgement. We are very grateful for the invaluable suggestions given by Ayush Kaushal. We also thank the organizers of the Shared Task.

References

1. Bhardwaj, M., Akhtar, M.S., Ekbal, A., Das, A., Chakraborty, T.: Hostility detection dataset in Hindi (2020). http://arxiv.org/abs/2011.03588
2. Chowdhury, S.A., Mubarak, H., Abdelali, A., Jung, S.g., Jansen, B.J., Salminen, J.: A multi-platform Arabic news comment dataset for offensive language detection. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6203–6212. European Language Resources Association, Marseille, France, May 2020. https://www.aclweb.org/anthology/2020.lrec-1.761
3. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, pp. 25–35. Association for Computational Linguistics, August 2019. https://doi.org/10.18653/v1/W19-3504, https://www.aclweb.org/anthology/W19-3504
4. Davidson, T., Warmsley, D., Macy, M.W., Weber, I.: Automated hate speech detection and the problem of offensive language. CoRR abs/1703.04009 (2017). http://arxiv.org/abs/1703.04009
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. 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), Minneapolis, Minnesota, pp. 4171–4186, June 2019. https://www.aclweb.org/anthology/N19-1423
6. Doiron, N.: https://huggingface.co/monsoon-nlp/hindi-bert

7. Hossain, M.Z., Rahman, M.A., Islam, M.S., Kar, S.: BanFakeNews: a dataset for detecting fake news in Bangla. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2862–2871. European Language Resources Association, Marseille, May 2020. https://www.aclweb.org/anthology/2020.lrec-1.349
8. Jain, K., Deshpande, A., Shridhar, K., Laumann, F., Dash, A.: Indic-transformers: an analysis of transformer language models for Indian languages (2020)
9. Jha, V.K., Hrudya, P., Vinu, P.N., Vijayan, V., Prabaharan, P.: DHOT-repository and classification of offensive tweets in the Hindi language. Procedia Comput. Sci. **171**, 2324–2333 (2020). http://www.sciencedirect.com/science/article/pii/S1877050920312448. Third International Conference on Computing and Network Communications (CoCoNet 2019)
10. Joshi, R., Goel, P., Joshi, R.: Deep learning for Hindi text classification: a comparison. In: Tiwary, U.S., Chaudhury, S. (eds.) IHCI 2019. LNCS, vol. 11886, pp. 94–101. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-44689-5_9
11. Kakwani, D., et al.: IndicNLPSuite: monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961, November 2020. https://www.aclweb.org/anthology/2020.findings-emnlp.445
12. Kaushal, A., Vaidhya, T.: Winners at W-NUT 2020 shared task-3: leveraging event specific and chunk span information for extracting COVID entities from tweets. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020) (2020). https://doi.org/10.18653/v1/2020.wnut-1.79
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2017)
14. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019). http://arxiv.org/abs/1907.11692
15. Malmasi, S., Zampieri, M.: Challenges in discriminating profanity from hate speech. CoRR abs/1803.05495 (2018). http://arxiv.org/abs/1803.05495
16. Mitrović, J., Handschuh, S.: upInf - offensive language detection in German tweets. In: Proceedings of the GermEval 2018 Workshop 14th Conference on Natural Language Processing, September 2018
17. Mittos, A., Zannettou, S., Blackburn, J., Cristofaro, E.D.: "And we will fight for our race!" A measurement study of genetic testing conversations on Reddit and 4chan. CoRR abs/1901.09735 (2019). http://arxiv.org/abs/1901.09735
18. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web (2016)
19. Ottoni, R., Cunha, E., Magno, G., Bernardina, P., Meira, W., Almeida, V.: Analyzing right-wing YouTube channels: hate, violence and discrimination (2018)
20. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. CoRR abs/1912.01703 (2019). http://arxiv.org/abs/1912.01703
21. Patwa, P., et al.: Overview of constraint 2021 shared tasks: detecting English COVID-19 fake news and Hindi hostile posts. In: Chakraborty, T., Shu, K., Bernard, R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021. CCIS, vol. 1402, pp. 42–53. Springer, Cham (2021)
22. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
23. Radford, A.: Improving language understanding by generative pre-training (2018)
24. Romero, M.: https://huggingface.co/mrm8488/HindiBERTa

25. Safi Samghabadi, N., Patwa, P., Srinivas, P.Y.K.L., 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, pp. 126–131. European Language Resources Association (ELRA), Marseille, France, May 2020. https://www.aclweb.org/anthology/2020.trac-1.20
26. Sreelakshmi, K., Premjith, B., Soman, K.: Detection of hate speech text in Hindi-English code-mixed data. Procedia Comput. Sci. **171**, 737–744 (2020). https://doi.org/10.1016/j.procs.2020.04.080, http://www.sciencedirect.com/science/article/pii/S1877050920310498. Third International Conference on Computing and Network Communications (CoCoNet 2019)
27. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and verification (2018). http://arxiv.org/abs/1803.05355
28. Waseem, Z., Davidson, T., Warmsley, D., Weber, I.: Understanding abuse: a typology of abusive language detection subtasks. In: Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, pp. 78–84. Association for Computational Linguistics, August 2017. https://doi.org/10.18653/v1/W17-3012, https://www.aclweb.org/anthology/W17-3012
29. Wijesiriwardene, T., et al.: ALONE: a dataset for toxic behavior among adolescents on Twitter. In: Aref, S., et al. (eds.) SocInfo 2020. LNCS, vol. 12467, pp. 427–439. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60975-7_31
30. Wolf, T., et al.: HuggingFace's transformers: state-of-the-art natural language processing. CoRR abs/1910.03771 (2019). http://arxiv.org/abs/1910.03771