Towards Analyzing the Efficacy of Multi-task Learning in Hate Speech Detection

Krishanu Maity[1](✉), Gokulapriyan Balaji[2], and Sriparna Saha[1]

[1] Department of Computer Science and Engineering, Indian Institute of Technology
Patna, Patna 801103, India
{krishanu_2021cs19,sriparna}@iitp.ac.in

[2] Indian Institute of Information and Technology, Design and Manufacturing,
Kancheepuram, Chennai, India

Abstract. Secretary-General António Guterres launched the United Nations Strategy and Plan of Action on Hate Speech in 2019, recognizing the alarming trend of increasing hate speech worldwide. Despite extensive research, benchmark datasets for hate speech detection remain limited in volume and vary in domain and annotation. In this paper, the following research objectives are deliberated (a) performance comparisons between multi-task models against single-task models; (b) performance study of different multi-task models (fully shared, shared-private) for hate speech detection, considering individual dataset as a separate task; (c) what is the effect of using different combinations of available existing datasets in the performance of multi-task settings? A total of six datasets that contain offensive and hate speech on the accounts of race, sex, and religion are considered for the above study. Our analysis suggests that a proper combination of datasets in a multi-task setting can overcome data scarcity and develop a unified framework.

Keywords: Hate Speech · Data scarcity · Single Task · Multi-Task

1 Introduction

Our world's communication patterns have changed dramatically due to the rise of social media platforms, and one of those changes is an increase in improper behaviors like the usage of hateful and offensive language in social media posts. On 15 March 2021, an independent United Nations human right expert said that social media has too often been used with "relative impunity" to spread hate, prejudice and violence against minorities[1]. Hate speech [15] is any communication that disparages a person or group on the basis of a characteristic such as color, gender, race, sexual orientation, ethnicity, nationality, religion, or other features. Hate speech detection is crucial in social media because it helps in ensuring a safe and inclusive online environment for all users. Even though social media platforms provide space for people to connect, share, and engage

[1] https://news.un.org/en/story/2021/03/1087412.

B. Luo et al. (Eds.): ICONIP 2023, LNCS 14452, pp. 317–328, 2024.
https://doi.org/10.1007/978-981-99-8076-5_23

with each other, the anonymity and ease of access to these platforms also make them attractive platforms for those who engage in hate speech.

Hate speech has serious consequences and can cause significant harm to its targets. It can lead to increased discrimination, bullying, and even physical violence. Moreover, it can contribute to the spread of misinformation, stoke fear and division, and undermine the fabric of society. The harm that hate speech causes is amplified in online spaces, where the reach and impact of messages can be much greater than in the real world. According to the Pew Research Center, 40% of social media users have experienced some sort of online harassment[2]. According to the FBI, there were 8,263 reported hate crime incidents in 2020, which represents an increase of almost 13% from the 7,314 incidents reported in 2019[3]. Between July and September 2021, Facebook detected and acted upon 22.3 million instances of hate speech content[4]. A study found that from December 2019 to March 2020, there was a substantial 900% surge in the number of tweets containing hate speech directed towards Chinese people and China[5]. These hate posts that are supposedly safe on social media create real-world violence and riots. This warrants the requirement for the detection and control of hate speech.

That is why social media companies have taken steps to detect and remove hate speech from their platforms. This is a challenging task, as hate speech often takes many different forms and is difficult to define. In addition, there is often a fine line between free speech and hate speech, and companies must balance these competing interests while still protecting users from harm. It is important to note that hate speech detection is not just a technical challenge, it is also a societal challenge. Companies must understand the cultural and historical context of hate speech to develop policies and algorithms that are fair and effective. It is also important to ensure that hate speech detection does not undermine freedom of expression, or discriminate against marginalized groups.

Over the last decade, plenty of research has been conducted to develop datasets and models for automatic online hate speech detection on social media [17,25]. The efficacy of hate speech detection systems is paramount because labeling a non-offensive post as hate speech denies a free citizen's right to express himself. Furthermore, most existing hate speech detection models capture only single type of hate speech, such as sexism or racism, or single demographics, such as people living in India, as they trained on a single dataset. Such types of learning negatively affect recall when classifying data that are not captured in the training examples. To build an effective machine learning or deep learning-based hate speech detection system, a considerable amount of labeled data is required. Although there are a few benchmark data sets, their sizes are often limited and they lack a standardized annotation methodology.

[2] https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

[3] https://www.fbi.gov/news/press-releases/fbi-releases-2019-hate-crime-statistics.

[4] https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/.

[5] https://l1ght.com/Toxicity_during_coronavirus_Report-L1ght.pdf.

In this work, we address three open research questions related to building a more generic model for textual hate speech detection.

(i) **RQ1:** *Does multi-task learning outperform single-task learning and single classification model trained using merged datasets?* This research question pertains to the advantage of multi-task learning for various datasets over other training strategies. When multiple datasets are available, the most intuitive method of training is to merge the datasets and train the model in a single-task learning setting. Different datasets are considered individual tasks in multi-task settings.

(ii) **RQ2:** *Which type of multi-task model performs the best across a wide range of benchmark datasets?* Two widely used multi-task frameworks, Fully shared (FS) and Shared private (SP) with adversarial training (Adv), have been explored to investigate which one is preferable for handling multiple datasets.

(iii) **RQ3:** *What combination of datasets improve or degrade the performance of the multi-task learning model?* This question addressed the effect of different dataset combinations on model performance. Different dataset combinations bring knowledge from various domains. For n datasets ($n >= 2$), there are ($2^n - n - 1$) possible combinations, each containing at least two datasets. The study on the improvement of performance on the grounds of complementary or contrasting properties of datasets plays an important role in the selection of datasets for multi-task learning.

This current paper addresses the above-mentioned questions by developing three multi-task learning models: fully shared, shared-private, and adversarial, as well as presenting insights about dataset combinations and investigating the performance improvement of multi-task learning over single-task learning and a single model trained using a merged dataset.

2 Related Work

Text mining and NLP paradigms have previously been used to examine a variety of topics related to hate speech detection, such as identifying online sexual predators, detecting internet abuse, and detecting cyberterrorism [22].

Detecting hateful and offensive speech presents challenges in understanding contextual nuances, addressing data bias, handling multilingual and code-switching text, adapting to the evolving nature of hate speech, dealing with subjectivity and ambiguity, countering evasion techniques, and considering ethical considerations [6]. These challenges necessitate robust and adaptable methodologies, including deep learning and user-centric approaches, to enhance hate speech detection systems. A common approach for hate speech detection involves combining feature extraction with classical machine learning algorithms. For instance, Dinakar et al. [3] utilized the Bag-of-Words (BoW) approach in conjunction with a Naïve Bayes and Support Vector Machines (SVMs) classifier. Deep Learning, which has demonstrated success in computer vision, pattern recognition, and speech processing, has also gained significant momentum in natural language processing (NLP). One significant advancement in this direction was the introduction of embeddings [14], which have proven to be useful when combined with classical machine learning algorithms for hate speech detection [13], surpassing the performance of the BoW approach. Furthermore, other Deep Learning methods have been explored, such as the utilization of Convolutional Neural Networks (CNNs) [27], Recurrent Neural Networks (RNNs) [4], and hybrid models combining the two [9]. Another significant development was the introduction of transformers, particularly BERT, which exhibited exceptional performance in a recent hate speech detection competition, with seven out of the top ten performing models in a subtask being based on BERT [26].

2.1 Works on Single Dataset

The work by Watanabe et al. [25] introduced an approach that utilized unigrams and patterns extracted from the training set to detect hate expressions on Twitter, achieving an accuracy of 87.4% in differentiating between hate and non-hate tweets. Similarly, Davidson et al. [2] collected tweets based on specific keywords and crowdsourced the labeling of hate, offensive, and non-hate tweets, developing a multi-class classifier for hate and offensive tweet detection. In a separate study, a dataset of 4500 YouTube comments was used by authors in [3] to investigate cyberbullying detection, with SVM and Naive Bayes classifiers achieving overall accuracies of 66.70% and 63% respectively. A Cyberbullying dataset was created from Formspring.me in a study by authors in [20], and a C4.5 decision tree algorithm with the Weka toolkit achieved an accuracy of 78.5%. CyberBERT, a BERT-based framework created by [17], exhibited cutting-edge performance on Twitter (16k posts), Wikipedia (100k posts) and Formspring (12k posts) datasets. On a hate speech dataset of 16K annotated tweets, Badjatiya et al [1] conducted extensive tests with deep learning architectures for learning semantic word embeddings, demonstrating that deep learning techniques beat char/word n-gram algorithms by 18% in terms of F1 score.

2.2 Works on Multiple Datasets

Talat et al. [23] experimented on three hate speech datasets with different annotation strategies to examine how multi-task learning mitigated the annotation bias problem. Authors in [21] employed a transfer learning technique to build a single representation of hate speech based on two independent hate speech datasets. Fortuna et al. [5] merged two hate speech datasets from different social media (one from Facebook and another from Twitter) and examined that adding data from a different social network allowed to enhance the results.

Although there are some attempts in building a generalized hate speech detection model based on multiple datasets, none of them has addressed the insight on (i) how to combine datasets; (ii) is multi-tasking better than single task setup and a single model trained using merged dataset, (iii) which type of multitasking is better: FS or SP.

Table 1. Source, statistics and domain of six hate speech datasets used in our experiments

Dataset	# Samples	# Classes and #Samples in each class	Source	Domain
D1 [2]	24783	3: Hate speech (1430), Offensive (19190), Neither (4163)	Twitter	Hate, Offensive
D2 [7]	10703	2: Non-hate (9507), Hate (1196)	Stormfront forum	Race, Religion
D3 [24]	10141	3: Racism (12), Sexism (2656), None (7473)	Twitter	Race, Sexism
D4 [12]	7005	2: Non Hate-Offensive (4456), Hate and Offensive (4456)	Twitter	Hate, Offensive
D5 [16]	10000	2: Non-hateful (5790), Hateful (4210)	Twitter	Immigrants, Sexism
D6 [11]	31962	2: Non-hate (29720), Hate (2242)	Twitter	Race, Sexism

3 Dataset Description

Six datasets (Table 1) are selected in an attempt to understand the effect of using multiple datasets and to conduct experiments. These datasets include examples of hate, offensiveness, racism, sexism, religion, and prejudice against immigrants. Even though the samples differ in terms of annotation style, domain, demography, and geography, there is common ground in terms of hate speech.

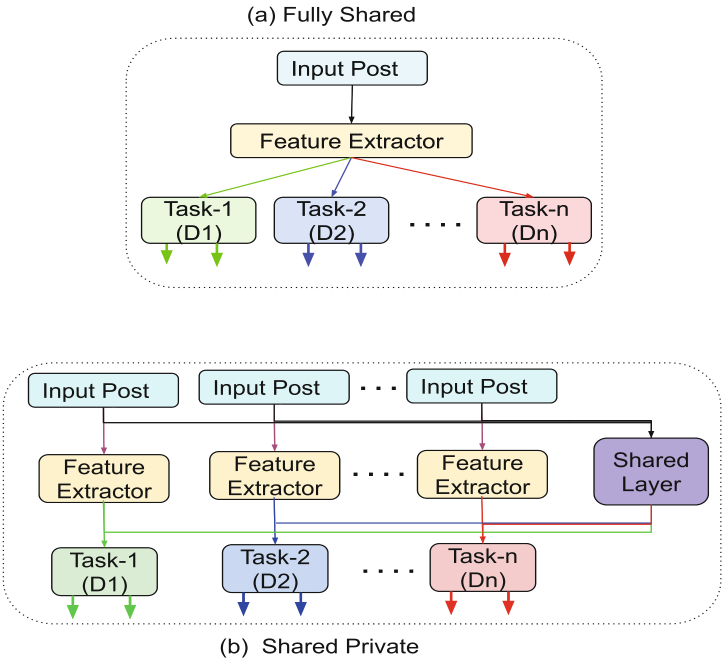
4 Methodology

To investigate how multiple hate speech datasets can help in building a more generalized hate speech detection model, we have experimented with two widely used multi-task frameworks (Fig. 1), i.e., Fully shared and Shared Private, developed by [10]. In the feature extraction module (Fig. 2), we employed Glove [18] and FastText [8] embedding to encode the noisy social media data efficiently. The joint embedding is passed through a convolution layer followed by max pooling to generate the local key phrase-based convoluted features. In the FS model, the final output from the CNN module is shared over n task-specific channels, one for each dataset (task). For the SP model, individual CNN representation from each of the tasks is passed through the corresponding task-specific output layer. In addition to task-specific layers, there is a shared layer (Fully Connected layer) to learn task invariant features for the SP model. The adversarial loss is added in model training to make shared and task-specific layers' feature spaces mutually exclusive [19].

5 Experimental Results and Analysis

This section describes the results of single task setting, multi-task setting of three models for different combinations of 6 benchmark datasets. The experiments are intended towards addressing the following research questions:

- **RQ1**: How does multi-task learning enhance the performance of hate speech detection compared to single task learning and single task based on a merged dataset?


(a) Fully Shared

(b) Shared Private

Fig. 1. (a) Fully shared and (b) Shared private multi-task frameworks.

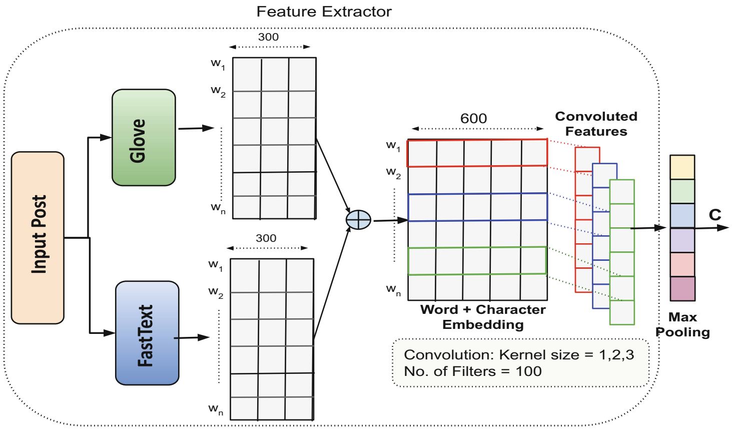

Feature Extractor

Fig. 2. Feature extraction module based on Glove and FastText joint embedding followed by CNN

- **RQ2**: Which type of multi-task learning model provides the best results among the three models?
- **RQ3**: Which combination of the benchmark datasets should be used for obtaining the best results from multi-task learning?

The experiments were performed on 5-fold cross-validation on the datasets and the results are evaluated in terms of accuracy value. The values mentioned inside the brackets are the improvements or decrements in accuracy compared to single-task learning. Keeping the size of the datasets in mind, a batch size of 8 was found optimal and configurations such as the ReLU activation function, and 5e−4 learning rate were chosen and the models were trained for 20 epochs.

Table 2. Single-task learning performance with individual datasets and merged datasets

Dataset Combination	Single Task		
	STL	Merged (All)	Merged (-D1)
D1	91.28	20.33	-
D2	87.6	84.96	88.97
D3	82.89	71.71	73.63
D4	63.81	63.74	64.88
D5	70.05	58.5	59.9
D6	94.75	87.63	92.87

Table 3. Multi-task Learning Performance

Dataset Combination	Multi Task			
	FS	FS - adv	SP	SP - adv
D1	92.68 (+1.40)	93.63 (+2.35)	95.04 (+3.76)	95.59 (+4.31)
D2	90.20 (+2.60)	89.02 (+1.42)	88.70 (+1.10)	89.53 (+1.93)
D3	83.81 (+1.12)	83.62 (+0.73)	86.79 (+3.90)	86.95 (+4.06)
D4	67.88 (+4.07)	66.25 (+2.44)	66.10 (+2.29)	65.53 (+1.72)
D5	71.45 (+1.40)	71.67 (+1.62)	74.80 (+4.75)	75.00 (+4.95)
D6	96.16 (+1.41)	95.72 (+0.97)	96.70 (+1.95)	96.78 (+2.03)

5.1 RQ1: Single Task vs Merging All vs Multi-task

In Table 2, the accuracy of single task learning is compared with a model trained after merging all datasets and with a multitasking framework. It is evident from this table that the performance of single-task learning is better than that of the model trained using a merged version of all the datasets. However, when dataset 1 which performed very poorly was removed from the merged set and experiments are again conducted, the accuracy values for datasets 2 and 4 are improved over the single-task learning accuracies. The selection of datasets that

Table 4. Experimental results of Fully Shared, Shared Private models under multi-task settings with 2 datasets combinations; Like, in (D3-D5) combination, 1st and 2nd represent the performance of D3 and D5, respectively

Dataset Combination	Fully Shared		Shared Private	
	1st	2nd	1st	2nd
D1-D2	93.33 (+2.05)	90.19 (+2.59)	94.05 (+2.77)	88.00 (+0.4)
D1-D3	93.55 (+2.27)	83.34 (+0.45)	94.01 (+2.73)	84.07 (+1.18)
D1-D4	93.54 (+2.26)	68.88 (+5.07)	93.93 (+2.65)	64.48 (+0.67)
D1-D5	93.35 (+2.07)	72.55 (+2.50)	93.40 (+2.12)	74.60 (+4.55)
D1-D6	92.39 (+1.11)	95.22 (+0.47)	94.61 (+3.33)	95.51 (+0.76)
D2-D3	89.86 (+2.26)	83.39 (+0.50)	89.37 (+1.77)	84.96 (+2.07)
D2-D4	90.55 (+2.95)	67.74 (+3.93)	88.27 (+0.67)	64.45 (+0.64)
D2-D5	90.00 (+2.4)	73.20 (+3.15)	89.25 (+1.65)	74.05 (+4.00)
D2-D6	90.43 (+2.83)	95.52 (+0.77)	88.46 (+0.86)	95.77 (+1.02)
D3-D4	83.88 (+0.99)	67.38 (+3.57)	84.22 (+1.33)	65.24 (+1.43)
D3-D5	83.00 (+0.11)	71.90 (+1.85)	84.57 (+1.68)	74.75 (+4.70)
D3-D6	83.44 (+0.55)	95.18 (+0.43)	84.17 (+1.28)	95.86 (+1.11)
D4-D5	68.09 (+4.28)	71.59 (+1.54)	65.31 (+1.50)	73.25 (+3.20)
D4-D6	67.09 (+3.28)	96.04 (+1.29)	65.42 (+1.61)	96.20 (+1.45)
D5-D6	72.05 (+2.00)	95.95 (+1.20)	73.80 (+3.75)	96.30 (+1.55)

are used to form the merged dataset for developing a unified model plays a significant role in the performance of the system. When the combination of datasets is selected after analyzing the domain, supplementary and complementary information available with the dataset, the unified model becomes more generalized. But blindly combining all the datasets leads to decreased performance of the unified model trained on the merged dataset. In multi-task settings (see Table 3), the performances on all the datasets are improved significantly over both single-task learning and single-task training on a merged dataset. In a multi-task setting, hate speech detection from a single dataset is considered an individual task. This concept proves to provide an edge to the model for its ability to generalize and perform better compared to the other training settings.

5.2 RQ2: Fully Shared vs. Shared Private (+/− Adversarial Training)

Among the models trained over multiple datasets as shown in Tables 4 and 5, there is no clear winner that can be selected. However, with the benchmark datasets used in our experiments, the shared private model proves to be the better model among its alternatives. This could be due to the training of shared and task-specific layers on the datasets which provide in-depth knowledge and prioritize the information from both these layers. But, the absence of such an ability to prioritize shared knowledge inhibits the performance of the fully shared network. As proof of this, the accuracies for datasets 1, 3, 5, and 6 among all the combinations are higher in the shared private model compared to the fully

Table 5. Experimental results of Fully Shared - Adversarial, Shared Private - Adversarial models under multi-task settings with 2 datasets combinations; Like, in (D3-D5) combination, 1st and 2nd represent the performance of D3 and D5, respectively

Dataset Combination	Fully Shared - Adversarial		Shared Private - Adversarial	
	1st	2nd	1st	2nd
D1-D2	93.51 (+2.23)	88.89 (+1.29)	94.69 (+3.41)	87.80 (+0.20)
D1-D3	93.67 (+2.39)	83.30 (+0.41)	94.96 (+3.68)	85.50 (+2.61)
D1-D4	93.60 (+2.32)	66.94 (+3.13)	94.67 (+3.39)	64.74 (+0.93)
D1-D5	93.28 (+2.00)	73.01 (+2.96)	94.71 (+3.43)	75.00 (+4.95)
D1-D6	92.30 (+1.02)	94.98 (+0.23)	94.39 (+3.11)	95.93 (+1.18)
D2-D3	89.95 (+2.35)	83.28 (+0.39)	88.51 (+0.91)	84.17 (+1.28)
D2-D4	90.03(+2.43)	66.87 (+3.06)	87.85 (+0.25)	64.54 (+0.73)
D2-D5	89.74 (+2.14)	73.24 (+3.19)	88.01 (+0.44)	72.85 (+2.80)
D2-D6	90.47 (+2.87)	95.47 (+0.72)	87.98 (+0.38)	95.91 (+1.16)
D3-D4	84.05 (+1.16)	66.83 (+3.02)	84.78 (+1.89)	64.77 (+0.96)
D3-D5	83.96 (+1.07)	72.11 (+2.06)	84.65 (+1.76)	74.98 (+4.93)
D3-D6	84.02 (+1.13)	95.50 (+0.75)	84.71 (+1.82)	95.95 (+1.20)
D4-D5	68.36 (+4.55)	71.52 (+1.47)	64.71 (+0.90)	73.92 (+3.87)
D4-D6	66.91 (+3.10)	95.83 (+1.08)	64.47 (+0.66)	96.66 (+1.91)
D5-D6	72.13 (+2.08)	95.98 (+1.23)	74.00 (+3.95)	96.45 (+1.70)

shared. However, interestingly the accuracy values of dataset 2 (D2) are better in a fully shared model. A possible explanation for this pattern could be in the source of the datasets. Unlike other datasets which were tweets, D2 belongs to a different source of social media posts.

When adversarial training is incorporated, the performance improves in datasets that have common ground/features. However, when the combination includes datasets of different sources, then the performance of the shared private adversarial model worsens compared to the shared private model. The adversarial layer alters the knowledge attained by the shared layer in such a way as to make the feature space of shared and specific layers to be mutually exclusive. This creates a more generalization causing deterioration in the performance. Fully shared adversarial is also similar in nature but the accuracy is hampered more compared to the shared private adversarial making this pattern difficult to predict or understand.

5.3 RQ3: Datasets Combination

From Table 6 and 7, it can be observed that the improvement in individual dataset compared to single task learning is limited as the number of datasets have increased (most of the time, the combination of two datasets performs better than the combination of three datasets). This could be due to the difficulty in generalizing the model on various datasets. The best performance is observed when using datasets of similar sizes and sources. An interesting insight was observed when datasets having information on different domains boost the

Table 6. Fully Shared Model Performance with 3 datasets combination

Dataset Combination	Fully Shared		
	1st	2nd	3rd
D1-D2-D3	92.27 (+0.99)	89.72 (+2.12)	83.44 (+0.55)
D1-D2-D4	92.25 (+0.97)	89.86 (+2.26)	68.31 (+4.50)
D1-D2-D6	92.21 (+0.93)	89.82 (+2.22)	95.06 (+0.31)
D1-D3-D4	92.35 (+1.07)	82.95 (+0.06)	68.74 (+4.93)
D1-D3-D5	91.97 (+0.69)	83.05 (+0.16)	71.15 (+1.10)
D1-D4-D5	91.83 (+0.55)	69.20 (+5.39)	70.95 (+0.90)
D2-D3-D5	90.05 (+2.45)	83.41 (+0.52)	71.60 (+1.55)
D2-D4-D6	90.01 (+2.41)	66.88 (+3.07)	95.17 (+0.42)
D3-D4-D5	83.40 (+0.51)	67.52 (+3.71)	71.15 (+1.10)
D4-D5-D6	67.38 (+3.57)	71.20 (+1.15)	94.90 (+0.15)

Table 7. Shared Private Model Performance with 3 datasets combination

Dataset Combination	Shared Private		
	1st	2nd	3rd
D1-D2-D3	94.67 (+3.39)	88.70 (+1.10)	84.33 (+1.44)
D1-D2-D4	94.57 (+3.29)	88.45 (+0.85)	65.02 (+1.21)
D1-D2-D6	94.59 (+3.31)	88.53 (+0.93)	95.02 (+0.27)
D1-D3-D4	94.45 (+3.17)	83.80 (+0.91)	64.64 (+0.83)
D1-D3-D5	95.05 (+3.77)	83.64 (+0.75)	72.24 (+2.19)
D1-D4-D5	94.49 (+3.21)	63.94 (+0.13)	72.67 (+2.62)
D2-D3-D5	88.78 (+1.18)	83.49 (+1.20)	72.22 (+2.17)
D2-D4-D6	88.51 (+0.91)	64.55 (+0.74)	95.77 (+1.02)
D3-D4-D5	84.05 (+1.16)	64.42 (+0.61)	73.43 (+3.38)
D4-D5-D6	64.67 (+0.86)	73.31 (+3.26)	95.88 (+1.13)

performance of each other significantly. For example, datasets 1 and 6 belonging to the same source have samples emphasizing different domains. Dataset 1 having samples that are majorly offensive gains shared knowledge on the attack of women and immigrants from dataset 6. Dataset 6 too learns knowledge of contrasting domains from dataset 1 that help generalize the model to tackle new samples.

6 Conclusion and Future Work

In this paper, an attempt was made to create a hate speech detection model that was trained on different datasets. To improve the performance and generality of the model, multi-task learning was leveraged. With the help of this methodology and careful examination of the datasets, a robust model that identifies and prevents various domains of hate attacks can be built, thus creating a safe and trustworthy space for users in social media. The contributions of the current work are twofold: (a) Experiments conducted across different types of settings and models help us develop a multi-task system that can be trained on datasets from different domains and detect hate speech in a generalized manner. (b) Studies were conducted on the effect of combinations and increase in datasets in a multi-task setting to improve the decision-making process of setting up new hate speech detection systems.

In the future, we would like to work on multi-modal hate speech detection systems that can help us monitor a plethora of social media.

Acknowledgements. The Authors would like to acknowledge the support of Ministry of Home Affairs (MHA), India, for conducting this research.

References

1. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759–760 (2017)
2. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, pp. 512–515 (2017)
3. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: 2011 Proceedings of the International Conference on Weblog and Social Media. Citeseer (2011)
4. Do, H.T.T., Huynh, H.D., Van Nguyen, K., Nguyen, N.L.T., Nguyen, A.G.T.: Hate speech detection on Vietnamese social media text using the bidirectional-LSTM model. arXiv preprint arXiv:1911.03648 (2019)
5. Fortuna, P., Bonavita, I., Nunes, S.: Merging datasets for hate speech classification in Italian. In: EVALITA@ CLiC-it (2018)
6. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (CSUR) **51**(4), 1–30 (2018)
7. de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate speech dataset from a white supremacy forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium, October 2018, pp. 11–20. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/W18-5102. https://www.aclweb.org/anthology/W18-5102
8. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893 (2018)
9. Maity, K., Saha, S.: BERT-capsule model for cyberbullying detection in code-mixed Indian languages. In: Métais, E., Meziane, F., Horacek, H., Kapetanios, E. (eds.) NLDB 2021. LNCS, vol. 12801, pp. 147–155. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-80599-9_13
10. Maity, K., Saha, S., Bhattacharyya, P.: Emoji, sentiment and emotion aided cyberbullying detection in Hinglish. IEEE Trans. Comput. Soc. Syst. **10**, 2411–2420 (2022)
11. Malik, J.S., Pang, G., van den Hengel, A.: Deep learning for hate speech detection: a comparative study. arXiv preprint arXiv:2202.09517 (2022)
12. Mandl, T., et al.: 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, pp. 14–17 (2019)

13. Mehdad, Y., Tetreault, J.: Do characters abuse more than words? In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 299–303 (2016)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
15. Nockleby, J.T.: Hate speech in context: the case of verbal threats. Buff. L. Rev. **42**, 653 (1994)
16. i Orts, Ò.G.: Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: frequency analysis interpolation for hate in speech detection. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 460–463 (2019)
17. Paul, S., Saha, S.: CyberBERT: BERT for cyberbullying identification. Multimed. Syst. **28**, 1897–1904 (2020)
18. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
19. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond binary labels: political ideology prediction of Twitter users. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 729–740 (2017)
20. Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to detect cyberbullying. In: 2011 10th International Conference on Machine Learning and Applications and Workshops, vol. 2, pp. 241–244. IEEE (2011)
21. Rizoiu, M.A., Wang, T., Ferraro, G., Suominen, H.: Transfer learning for hate speech detection in social media. arXiv preprint arXiv:1906.03829 (2019)
22. Simanjuntak, D.A., Ipung, H.P., Nugroho, A.S., et al.: Text classification techniques used to faciliate cyber terrorism investigation. In: 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies, pp. 198–200. IEEE (2010)
23. Talat, Z., Thorne, J., Bingel, J.: Bridging the gaps: multi task learning for domain transfer of hate speech detection. In: Golbeck, J. (ed.) Online Harassment. HIS, pp. 29–55. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78583-7_3
24. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL Student Research Workshop, San Diego, California, June 2016, pp. 88–93. Association for Computational Linguistics (2016). http://www.aclweb.org/anthology/N16-2013
25. Watanabe, H., Bouazizi, M., Ohtsuki, T.: Hate speech on Twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access **6**, 13825–13835 (2018)
26. 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)
27. Zimmerman, S., Kruschwitz, U., Fox, C.: Improving hate speech detection with deep learning ensembles. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018 (2018)