Lifelong Learning Framework for Multilingual Hate Speech Detection in Social Media Streams

Yasas Senarath[0000−0002−8759−378X] Hemant Purohit[0000−0002−4573−8450]

George Mason University, Fairfax, VA, USA
ywijesu@gmu.edu, hpurohit@gmu.edu

Abstract. Online hate speech poses a threat to the harmony of both online and real-world communities. Due to the proliferation of such speech, advancements in the research on hate speech detection have been growing in recent years. However, we note that there is a research gap on multilingual hate speech detection in online social media streams. In this paper, we showcase extensive evaluations of multilingual hate speech detection within a lifelong machine learning framework. We assess state-of-the-art techniques in lifelong machine learning for the task of hate speech detection and critically analyze the performance of our approach against strong baselines.

Keywords: Multilingual Hate Speech Detection · Lifelong Machine Learning · Deep Learning.

1 Introduction

Online hate speech has become a pressing issue in the digital age. Hate speech can have severe consequences on society, such as inciting violence, decreasing social cohesion, and promoting discrimination. A community with low social cohesion is more likely to have less community resilience, which is an essential factor in recovering from a crisis. Therefore it is important to mitigate the spread of hate speech in online platforms. Detecting hate speech in real-time is one of the core steps of mitigating hate speech.

Hate speech detection is a challenging task due to the dynamic nature of the language dialect (form) used in online platforms. The dialect used in online platforms is constantly evolving and can vary across different topics of discussion. For example, hate speech against African American women may take a different form than hate speech against Caucasian women. Therefore it is challenging to use a single static model model to detect hate speech across these different topics in real time. However, the research in hate speech detection has been focused on a single topic or a limited set of topics that were present in the training dataset. In real-world scenarios, the topics of online hate speech change over time due to the social, political, and economic events happening in the world. Therefore, it is important to develop a model that can adapt to new topics over time. This could be achieved by developing a lifelong learning model for hate speech detection.

Table 1. Example of Hate Speech in Different Languages

Topic	Language	text
Immigration	English	Outrageous!!!!!!!! Outrageous!!! youre going illegal aliens
Immigration	French	c'est un golden retriever une race créée par des blancs ces ching chong ne méritent ni nos chiens ni nos inventions. (it's a golden retriever a breed created by white people these ching chong deserve neither our dogs nor our inventions.)

Another challenge in hate speech detection is the multilingualism of online platforms (see Table 1 for examples with the same topic but different languages). Users of social media platforms post content in various languages. Therefore, it is important to develop models that can detect hate speech in many languages. However, the research that has been done in hate speech detection has been mostly focused on a single language or a limited set of languages without proposals for model adaptation to support new languages in a lifelong manner.

In this paper, we propose the problem of lifelong hate speech detection for multilingual hate speech detection. Then propose learning approaches based on state-of-the-art deep learning techniques used in the hate speech detection and lifelong learning literature. We evaluate the proposed approach on a bilingual dataset and show that the proposed approach outperforms the baselines. We use established techniques used in modeling lifelong text classification and hate speech detection and apply them to our problem of multilingual lifelong hate speech detection.

Lifelong learning is a machine learning paradigm that aims to develop a system that can learn continuously from a stream of "tasks" in a human-like learning approach. In lifelong learning, a model learns on a sequence of tasks and uses the knowledge gained from previous tasks to improve the performance on newer tasks while maintaining a consistent performance on previous tasks. We define a single task as a binary classification of labels (hate speech and normal speech) on a specific topic in a specific language. Since this sort of experiment has not been performed previously we simulate the task stream by dividing an existing multilingual dataset into tasks based on the topics and languages of the document. The topic of the document is inferred by the topic distribution of the document from topic modeling. This has been illustrated in Figure 1. At each step of the task stream, we assume that there is a labeled dataset to train the model. We evaluate a model on the current task and all past tasks after training the current task.

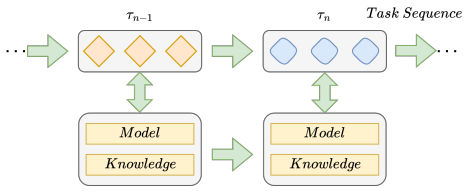

Fig. 1. Online data stream scenario for lifelong hate speech detection

The rest of the paper is organized as follows. In Section 2, we discuss the related work in hate speech detection and lifelong machine learning. We discuss the methodology used in the experiments in Section 3. In Section 4 and Section 5, we discuss the experimental setup and the results of the experiments accordingly. Section 6 concludes the paper.

2 Related Work

2.1 Hate Speech Detection

Detection of hate speech and its various forms in online platforms has been extensively studied in the literature. Early works on hate speech detection focused on developing traditional machine learning models for hate speech detection. [6] proposed a bag-of-words (BOW) supervised machine learning classification approach to identify cyberbullying in social networks. [19] used topic models to automate feature generation to help with training a hate speech detection model in a semi-supervised manner using Logistic Regression. Recently, there has been growing interest in using deep learning models for hate speech detection [1,9,17,21,8]. In the study by [14], the authors have showcased the importance of character level features on hate speech detection using hybrid of traditional machine (Linear Regression and Support Vector Machine) and deep learning methods (Convolutional Neural Network). [18] proposed an identity based framework for exploring the possibility of leveraging an identity based framework for generalizable hate speech detection since it has been identified that hate speech detection models are biased based on the identity of the target group [2,20].

Multilingual Hate Speech Detection: [15] proposed a multilingual multi-aspect hate speech dataset and evaluated the performance of state-of-the-art models at the time of study on this dataset and found that deep learning models perform better than traditional BOW-based models. Furthermore, [5] evaluated the performance of common deep learning configurations and found that word embedding-backed Long Short Term Memory (LSTM) models outperform other deep learning methods in multilingual hate speech detection.

2.2 Lifelong Machine Learning

Lifelong learning as defined in [7,11] refers to the learning paradigm where a model learns on a sequence of tasks incrementally, accumulates the learned knowledge, and uses it to help future learning while updating the knowledge as required. [4] proposed a lifelong learning approach to sentiment classification of reviews. They accumulated token-level knowledge on tasks defined based on domain and used them for learning as additional loss in the stochastic gradient descent (SGD) algorithm. Moreover, [16] proposed a lifelong learning approach for hate speech detection where they used a memory module based on LB-SOINN (Load-Balancing Self-Organizing Incremental Neural Network) to retain embedding representations of important instances from previous tasks. They used these

instances when updating the model on newer tasks. We use the same approach in our work to develop a lifelong hate speech detection model for multilingual hate speech detection.

3 Methodology

3.1 Problem Formulation

Let a sequence of n tasks be $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$, where each task τ_i is a binary hate speech detection task. The goal of the (supervised) lifelong learning system is to learn a model that can perform well on all tasks in the sequence. It should maintain a consistent performance on previous tasks as new tasks are introduced. Moreover, when the model is trained on new tasks, it should use the knowledge from previous tasks to improve the performance on the new tasks. In online platforms, hate speech discloses itself in different topics and languages and the performance of a model can degrade upon the introduction of a new topic or language. In this work, we define a task to accommodate the change in topic or the language of the message. Therefore, the task is defined as detecting hate speech on a specific topic (t) in a specific language l and re-represented as $\tau_{(t,l)}$. Now, once we have defined the task, we can define the supervised lifelong hate speech detection problem as follows:

Problem (Multilingual Lifelong Hate Speech Detection): A learner has performed a sequence of supervised hate speech detection tasks from 1 to $n-1$ where each task has a training dataset labeled with classes hate speech and non-hate speech (normal). Given a new task τ_n it uses the knowledge gained in the past tasks to learn a better classifier for the new task.

We acknowledge that, in real-world scenarios, the same tasks may not appear simultaneously for different languages. However, for the sake of simplicity in the experiment simulation, we assume that tasks appear in the same order for each language and that the same task for each language appears at the same time.

In such a scenario, we have a system that gets trained over time (task stream) with the assumption that the topics of the tasks appear in the same order for each language and that all languages appear together for a given topic. The model is evaluated on all past tasks as well as the current task. This ordering of task sequence is illustrated in Figure 2.

3.2 Lifelong Multilingual Hate Speech Detection

The proposed lifelong multilingual hate speech detection system first receives a task from the task stream. It also expects that human annotated data is available for the task. The system then uses the learning component to improve its detection capabilities. In addition to the data provided for the learning, it will store any instances from past tasks that are saved in the memory. The system architecture is illustrated in Figure 3. There are two major components in the system:

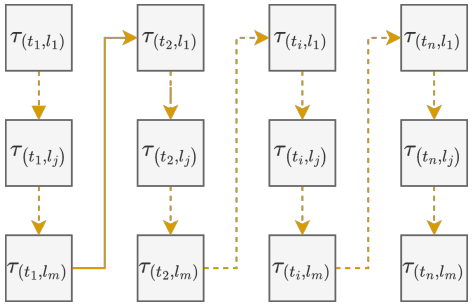

Fig. 2. Simulated Task Stream for Multilingual Lifelong Hate Speech Detection

1) Learning component: We use a pre-trained multilingual language model as the base sentence encoder and use stochastic gradient descent (SGD) to fine-tune on the current task with the help of the AdamW optimizer [12]. Pre-trained language models have shown exceptional performance on text classification tasks after fine-tuning on those provided tasks. When learning a new task such as hate speech detection, these models can leverage the knowledge it has learned from the large corpus of text when the model was pre-trained. We use a language model as an encoder and add a linear layer on top of the encoder to classify the text as hate speech or normal speech. Additionally, we use two loss components in the SGD algorithm: the cross-entropy loss and memory based loss.

2) Memory component: We keep track of the most important instances of a task after training on a given task to be used in the future when training on newer tasks. To identify the most important instances, we use the Load-Balancing Self-Organizing Incremental Neural Network (LB-SOINN) similar to [16]. We refer the reader to [16] for more details on the memory component. While training the model on a new task, we use the instances stored in the memory component and compute the loss on these instances as an additional loss in the SGD algorithm to prevent catastrophic forgetting.

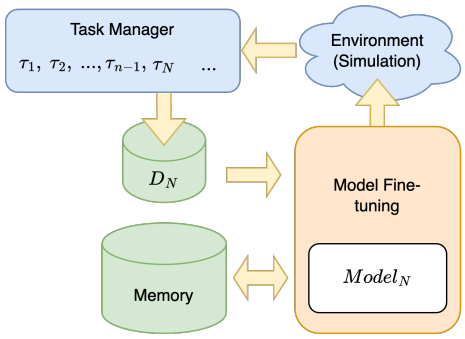

Fig. 3. Architecture of Proposed Multilingual Lifelong Hate Speech Detection System

4 Experimental Setup and Evaluation

4.1 Dataset

In this study, we only utilize the dataset from [15]. The dataset consists of hate speech tweet text data in English, French, and Arabic languages. However, in this study, we only use the English and French data. This dataset is annotated with multiple labels: Abusive, Hateful, Offensive, Disrespectful, Fearful and Normal. We label tweets as 'positive' for hateful content if they are not marked as normal; otherwise, they are labeled as 'negative' for hateful content. This binary labeling approach follows the same methodology as [15] in binarizing the labels. In the experiments, we will use these binary labels as the ground truth.

Preprocessing. The dataset was already preprocessed by the authors of the dataset. We use the preprocessed dataset for the experiments. The usernames and URLs were replaced with @user and @url respectively. Moreover, we augment the dataset by translating the French tweets to English using the models provided by the HuggingFace Transformers library. These translated tweets serve as additional data in our experiments, ensuring a balanced representation for each language across all topics, as the original dataset lacked sufficient data for equivalent topics in both languages.

4.2 Task Stream

In the problem formulation section, we have defined a task as detecting hate speech on a specific topic in a specific language. We use this definition to create a task stream for the experiments. Due to the absence of real-world data stream access and timestamps for individual posts, we simulate the task stream by partitioning the dataset into tasks based on document topics and languages. We have used the topic distribution of the documents from topic modeling to assign a topic to each document. We then group the documents by topics and languages and arrange them in a sequence based on the number of documents in each group to simulate a data stream. We use this data stream to train the model in a lifelong learning scenario.

Topic Modeling: We identify the topic of an instance by using topic modeling. Specifically, we use the approach proposed by BERTopic [10]. This topic model is based on sentence embeddings and clustering to identify the topics. We use BERTopic with *distiluse-base-multilingual-cased-v1* sentence encoder, UMAP [13] dimensionality reduction, and HDBSCAN [3] clustering to build the topic model on a subset of the dataset. We kept the number of topics to 8 to minimize over-clustering and get a good representation of the topics in the dataset. The number of topics was manually calibrated such that we have a reasonable number of training examples to train the models. Furthermore, we have set the minimum number of documents in a topic to 50 to avoid topics with very few documents. By this method, we were able to identify semantically similar messages in the text that we assumed would appear together in the task stream.

4.3 Performance Measures and Approach

Since the test data is imbalanced, we use the F1 score as the primary performance measure. We calculate the F1 scores of the current task and all past tasks after training on the new (current) task. Therefore, we get a series of F1 scores for each task in the task stream indicating the performance of the model after being adapted on a newer task.

Table 2. Average F1 Scores of Models on Test Data of Each Task

Task ID	1	2	3	4	5	6	7	8	9	10
Batch Learning	0.47	0.51	0.49	0.44	0.43	**0.44**	**0.44**	0.55	0.52	**0.63**
Online Learning	0.46	0.54	0.46	0.35	0.41	0.37	0.38	0.34	**0.53**	0.44
Lifelong Learning	**0.53**	**0.63**	**0.60**	**0.52**	**0.45**	**0.44**	0.06	**0.59**	0.50	0.55

4.4 Experimental Setup

We use the same learning hyperparameters across all the baseline and proposed models. We have set the learning rate to $1e-5$, batch size to 8, and the number of epochs to 10. However, we also use early stopping with a patience of 3 to prevent overfitting.

We compare our proposed method to two other baseline methods. **Baseline 1: Batch Learning** is a model trained on each task independently without any knowledge transfer across tasks. **Baseline 2: Online Learning** is an online learning model without any explicit method to prevent catastrophic forgetting. We evaluate the performance of the proposed model and the baselines on the task stream and compare the results.

5 Results

There are two major results that we present in this section. First, we compare the evaluation results of the different approaches across tasks in the task stream. Second, we look at the comparison of the performance of the approaches solely on the new task at each task step.

5.1 Performance Across Tasks

Figure 4 shows the performance of the different approaches across tasks in the task stream. We observe that the proposed approach with lifelong learning has consistently been able to maintain a good performance across tasks. It is also confirmed by the average F1 scores in Table 2 where there is a clear improvement in the performance of the proposed model compared to the baselines (7 out of 10 tasks). This could be attributed to the memory component that helps the model

retain important instances from past tasks and use them to prevent catastrophic forgetting. On the other hand, the baselines have shown a drop in performance on the new tasks (especially towards the end). This is expected as the baselines do not have any mechanism to prevent catastrophic forgetting. Another important observation is the outlier tasks (task 7). We believe it to be due to the nature of topic being discussed in the task. We note however that the performance of task 7 increases after training on task 8 (same topic but different language). This is an interesting observation that we will investigate further in future work.

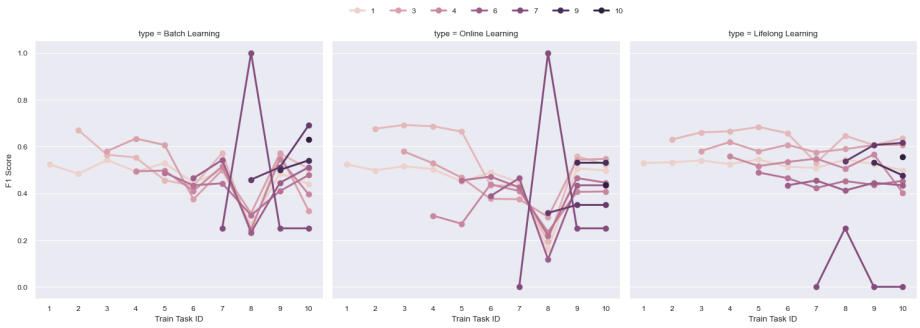

Fig. 4. Performance of the Proposed Model Across Tasks in the Task Stream

5.2 Performance on New Task

Another important aspect is to just look at the performance of the models on the new task at each step. This is illustrated in Figure 5. We observe that the proposed lifelong learning (Average F1=0.48) method has consistently outperformed the online learning approach (Average F1=0.42). However, the batch learning (Average F1=0.51) approach has shown a better performance on the new task at each step. We believe this is expected as in the batch learning process we train a model exclusively for that given task.

6 Conclusion

This paper introduced a novel method for the lifelong multilingual hate speech detection problem, by combining state-of-the-art techniques used in lifelong machine learning and multilingual hate speech detection. Our proposed approach utilized a pre-trained multilingual language model as the base sentence encoder and included a memory component to prevent catastrophic forgetting. We evaluated this approach on a bilingual dataset of social media posts using a simulated data stream setup and compared the results with two baselines. The results demonstrated that the lifelong learning based approach outperformed

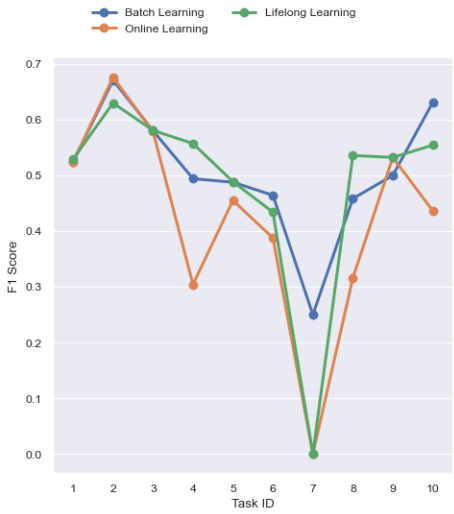

Fig. 5. Performance on the New Task at Each Task Step

the baselines in terms of F1 score. Future work should focus on extending the proposed methods to encompass a broader range of languages and topics and should explore semi-supervised learning techniques to further enhance the proposed approach, enabling more effective adaptation to continuous streams of largely unlabeled data.

Acknowledgments

Authors acknowledge the partial support of research grants from the Western Norway Research Institute's SOCYTI project, grant # 331736, and INTPART DTRF project, grant # 309448.

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. Borkan, D., Dixon, L., Sorensen, J., Thain, N., Vasserman, L.: Nuanced metrics for measuring unintended bias with real data for text classification. In: Companion proceedings of the 2019 world wide web conference. pp. 491–500 (2019)
3. Campello, R.J., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Pacific-Asia conference on knowledge discovery and data mining. pp. 160–172. Springer (2013)
4. Chen, Z., Ma, N., Liu, B.: Lifelong learning for sentiment classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 750–756 (2015)

5. Corazza, M., Menini, S., Cabrio, E., Tonelli, S., Villata, S.: A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology (TOIT) **20**(2), 1–22 (2020)
6. Dinakar, K., Jones, B., Havasi, C., Lieberman, H., Picard, R.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS) **2**(3), 1–30 (2012)
7. Fei, G., Wang, S., Liu, B.: Learning cumulatively to become more knowledgeable. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1565–1574 (2016)
8. Founta, A.M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., Leontiadis, I.: A unified deep learning architecture for abuse detection. In: Proceedings of the 10th ACM conference on web science. pp. 105–114 (2019)
9. Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the first workshop on abusive language online. pp. 85–90 (2017)
10. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)
11. Liu, B.: Learning on the job: Online lifelong and continual learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13544–13549 (2020)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017), https://api.semanticscholar.org/CorpusID:53592270
13. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software **3**(29) (2018)
14. 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)
15. Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., Yeung, D.Y.: Multilingual and multi-aspect hate speech analysis. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4675–4684 (2019)
16. Qian, J., Wang, H., ElSherief, M., Yan, X.: Lifelong learning of hate speech classification on social media. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2304–2314 (2021)
17. Singh, V., Varshney, A., Akhtar, S.S., Vijay, D., Shrivastava, M.: Aggression detection on social media text using deep neural networks. In: Proceedings of the 2nd workshop on abusive language online (ALW2). pp. 43–50 (2018)
18. Uyheng, J., Carley, K.M.: An identity-based framework for generalizable hate speech detection. In: International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. pp. 121–130. Springer (2021)
19. Xiang, G., Fan, B., Wang, L., Hong, J., Rose, C.: Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 1980–1984 (2012)
20. Yoder, M.M., Ng, L.H.X., Brown, D.W., Carley, K.M.: How hate speech varies by target identity: A computational analysis. arXiv preprint arXiv:2210.10839 (2022)
21. Ziqi, Z., Robinson, D., Jonathan, T.: Hate speech detection using a convolution-lstm based deep neural network. IJCCS **11816**, 2546–2553 (2019)