Efficient Detection of Multilingual Hate Speech by Using Interactive Attention Network with Minimal Human Feedback

Fedor Vitiugin
fedor.vitiugin@upf.edu
Universitat Pompeu Fabra
Barcelona, Spain

Yasas Senarath
ysenarath@gmu.edu
George Mason University
Fairfax, USA

Hemant Purohit
hpurohit@gmu.edu
George Mason University
Fairfax, USA

ABSTRACT

Online hate speech on social media has become a critical problem for social network services that has been further fueled by the self-isolation in the COVID-2019 pandemic. Current studies have primarily focused on detecting hate speech in one language due to the complexity of the task; however, hate speech has no boundaries across the languages and geographies in the real world nowadays. This demands further investigation on multilingual hate speech detection methods, with strong requirements for model interpretability to effectively understand the context of the model errors. In this paper, we propose a Multilingual Interactive Attention Network (MLIAN) model for hate speech detection on multilingual social media text corpora, by building upon the attention networks for interpretability and human-in-the-loop paradigm for model adaptability. This model interactively learns to give attention to the relevant contextual words and leverage the labels for the hate target mentions from the simulated human feedback. We evaluated the proposed model on SemEval-2019 Task 5 datasets in English and Spanish. Extensive experimentation of model training in both settings of single and multiple language data demonstrates the superior performance of our model (with AUC more than 84%) compared to the strong baselines. Our results show that human feedback not only improves the model performance but also helps to improve the interpretability of the model by establishing a strong connection between the learned attention weights and semantic frames for the text across languages. Further, an analysis of the amount of human feedback required to achieve reliable and increased model performance shows that less than 4% of training data is sufficient. The application of the MLIAN method can inform future studies on multilingual hate speech.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

Hate Speech Detection, Social Media, Human-in-the-loop Machine Learning

WebSci '21, June 21–25, 2021, Virtual Event, United Kingdom

ACM ISBN 978-1-4503-8330-1/21/06…$15.00
https://doi.org/10.1145/3447535.3462495

ACM Reference Format:
Fedor Vitiugin, Yasas Senarath, and Hemant Purohit. 2021. Efficient Detection of Multilingual Hate Speech by Using Interactive Attention Network with Minimal Human Feedback. In *13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447535.3462495

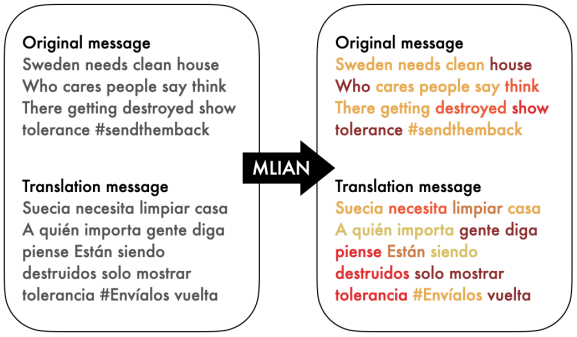

Figure 1: Attention weights distribution in English and Spanish texts. The darker color of the word means the higher weight.

1 INTRODUCTION

Hate speech has become a general phenomenon in modern society. Particularly, the prevalence of hate speech is pronounced on social media platforms and other means of online communication. Users often anonymously and freely express themselves in online communication forums, including social media. The ability to express freely oneself is an important human right, but inducing and spreading hate towards another group is an abuse of this liberty. While the research in hate speech detection has been growing rapidly, multilingual hate speech detection is still a challenging task. Most of the existing research studies [5, 29, 33] have focused on one language only (mainly English), and their methods often depend on external knowledge sources, such as a hate speech lexicon [10, 22, 41]. These sources are resource-intensive and time-consuming to create in every language. A key challenge of these methods is the obsolescence of the data source, given the developing language of user-generated content on social media. The language changes quickly, especially under the pressure of moderation, which brings us to the next important challenge. Current multilingual hate speech detection models cannot effectively deal with the local derogatory slangs in specific language (e.g., 'sudaca' is xenophobic term to call people from South America by Central Americans and North Americans who speak Spanish language) and local context of implicit hate

speech (e.g., 'building a wall' could implicitly refer to hate against immigrants). Another enormous challenge is a significant drop in the performance when the existing models are tested on datasets different from training data [22] in terms of the target of hate speech such as *immigrants* versus *women*. Examples of such tweets are presented in Table 1.

Table 1: Example of messages with different targets of hate speech.

Message	Target	
The U.S must stop importing the Worlds Poor if they cant take care of themselves #sendthemback Stop allowing Foreigners to live off U.S Taxpayers #Trump #MAGA	Immigrants	Group
@USER @USER You won the "life time recipient for Hysterical Woman" a long time ago	Woman	Individual

Last, the state-of-the-art hate speech detection models have used deep learning techniques, however, the decisions made by the deep learning models can be opaque and difficult for humans to interpret why the decision was made and analyze the model errors. While human-in-the-loop paradigm has been shown to assist such techniques, there is still a challenge for the ability of humans to provide effective feedback to the model to improve it. To address this challenge, we hypothesize that a theoretical approach of *frame semantics* from cognitive linguistics [17] can help better explain and rectify the model reasoning provided through attention weight map (c.f. Figure 1). Frame semantics suggests that word meanings are defined relative to frames in a given text and thus, if the model can learn to give attention to the elements of the correct frame as per human interpretation, the model performance could improve. For example, in Table 1, in order to correctly interpret the posts, a model will need a good understanding of the targets of the hate that could be easily understood by evoking a specific frame for interpretation by a human when looking over the attention maps.

This paper investigates the following research questions:

- RQ1. Can a hybrid method of Interactive Attention Network (IAN) with human-in-the-loop approach improve the detection of hate speech in multilingual data with local slangs and implicit context for hate?
- RQ2. Is there an effect of framing in the human feedback to IAN that helps toward faster convergence for the hate detection model?
- RQ3. How much human feedback is required for significant improvement of hate speech detection results?

To address these questions, our method relies on including minimal human guidance in the training process of IAN classification model for achieving higher performance. Human feedback helps to detect hate subtleties and phrases for extracting features during model training, where the human feedback is guided by common element of the frames to express a hate speech, i.e. hate targets. Explanation of decisions of the IAN model with the help of human feedback is analyzed by comparing the distribution of attention weight maps and semantic frames in the textual posts. The experimental dataset includes social media posts from two different topics in two languages. The proposed method allows the design of a novel multilingual hate speech detection system with the help of humans that shows high level of performance and can explain decisions by demonstrating an attention weight map (c.f. Figure 1) of the analyzed texts.

The main contribution of this study is a Multilingual Interactive Attention Network (MLIAN) model for detecting hate speech in text, regardless of language. We not only show improved model performance compared to baselines in two languages but also show a principled way of integrating frame semantics for analyzing the interpretability of the model reasoning. A comparison of distributed attention weight map with semantic frames shows that our model accurately captures implicit frame elements in the text that help to detect hate speech. We achieved this with simulated human feedback and identified the minimal level of feedback required to improve the model performance compared with the baselines.

The remainder of this paper is organized as follows. We first describe the related work, followed by our MLIAN methodology, experimental setup, and then, result analyses.

2 RELATED WORK

Spreading hate towards distinct groups is an abuse of human rights to express themselves freely. Many online forums such as Facebook and Twitter have policies to remove hate speech content [16], albeit detecting hate speech is challenging. We summarize the definitions, existing detection techniques, and the role of human feedback to improve them.

2.1 Hate Speech Definitions

There are many definitions of hate speech that make the task specification of the detection of hate speech difficult. Here are some examples of such definitions: (1) "Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity." [27] (2) "We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation." [16] (3) "Language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." [10] (4) "Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used." [18] Some definitions above consider hate towards a group while others consider attacks on an individual. A general observation among the definitions is that some aspect of the group's or a person's identity becomes a base for the offense, however, in the given text it may not be explicitly stated, as shown in Table 1. While in one definition the specific identity aspect is

ignored, other definitions provide specific identity characteristics. These ambiguities can challenge the task specification and what conventional text-based classification approaches could capture, especially when data is multilingual. Thus, human feedback could be valuable to support model training process.

2.2 Hate Speech Detection Models

The earlier efforts to build hate speech classifiers used both simple methods using dictionary lookup [23] and bag-of-words features [8] as well as deep learning techniques. SVM classifier is widely used for hate speech detection. Training includes diverse features such as character n-grams, word n-grams, word skip-grams, and knowledge-base features [30, 39]. Also, the list of features can include Brown cluster features along with approaches including ensemble classifiers and meta-classifiers [31]. During the analysis of the TRAC-1 workshop results [26], we found that authors use both deep learning (e.g., LSTM, Bi-LSTM, CNN) and traditional machine learning classifiers (e.g., SVM, Random Forest, Naive Bayes).

The HASOC 1 workshop organized at FIRE 2019 [32] notes that the most widely used approach for hate speech detection in Indo-European Languages was Long Short-Term Memory (LSTM) networks coupled with word embeddings. The participants used a wide variety of models such as BERT, SVM, and LSTM. Furthermore, a unified deep learning architecture based on LSTM networks reached high performance without change of the architecture but only training a model for each task (i.e., different abusive behavior types) [19]. Results of the most recent shared tasks in aggression identification and misogynistic aggression identification show that the superior performance of the SVM classifier was achieved mainly because of its better prediction of the majority class. BERT-based classifiers were found to predict the minority classes better [6].

Research into the multilingual aspect of hate speech is relatively new. Using Twitter hate speech corpus from five languages annotated with demographic information, authors of [24] studied the demographic bias in hate speech classification. Hate speech detection models based on SVM and Bi-LSTM show outstanding performance on three datasets from three languages (English, Italian, and German) [9]. Moreover, large-scale analysis of deep learning models to develop classifiers for multilingual hate speech classification (16 datasets from 9 languages) shows that for low resource languages, LASER embedding with logistic regression performs the best, while in a high resource setting BERT-based models perform better [3]. One limitation of these methods is the lack of interpretability of the reasoning for the models' decisions.

2.3 Human-Machine Collaboration for Hate Speech Detection

Complex behaviors such as hate speech require efficient automated models for detection of such communication. Previous work in this direction informs the requirement for continuous model transformation techniques [40] due to the complexity of the task. The findings of human-machine collaboration for content regulation (based on Reddit case) suggest a need for tools to help tune the performance of automated modeling mechanisms, a repository for sharing tools, and improving the division of labor between human and machine decision making [25]. Some existing approaches advocate the use of external knowledge sources, such as a hate speech lexicon, where detection systems could leverage multilingual, fine-grained Profanity and Offensive Word (POW) lexicons (an NLP resource for toxic language) [12]. This type of approach can be effective but it requires developing these knowledge sources that is labor-intensive, especially for multilingual setting, and furthermore, such sources need to be up to date, which is not always possible. Thus, an effective alternative can be a human-machine collaboration to fine-tune an automated hate speech detection model during the training/re-training process, with targeted human feedback to improve the model's understanding for the hate speech context.

3 METHODOLOGY: MLIAN MODEL

Recent approaches for hate speech detection propose solutions that use deep learning techniques for text classification of hate speech. While these solutions make decisions automatically, they make errors due to the biases in learning patterns and the reasoning behind those decisions can be difficult to interpret and unclear for humans. The resulting systems that automatically censor social media posts would end up needing a human's attention for majority of the appealed cases. The multilingual content can make such systems even more human resource-demanding. In this section we describe our proposed MLIAN model that can enable efficient human-in-the-loop paradigm along with interpretability of multilingual hate speech classification decisions, by employing a meaningful human feedback guided through frame semantics theory in the deep learning architecture of interactive attention networks.

3.1 Interactive Attention Networks

Our method builds upon the interactive attention network (IAN) architecture. Deep learning models are widely used for hate speech detection tasks [5], but many of such models make automated decisions hard for understanding, or can be explainable only by using special techniques [29]. In recent years, models with attention mechanism have not only shown good performance, but also can be used as a tool for interpreting the behavior of neural network architectures [11, 20, 21]. In the processing of natural language, the tokens composing the source text are characterized by having each a different relevance to the task at hand. The attention mechanism constructs the context vectors of the tokens that are required by the decoder to generate the output sequence in the encoder-decoder neural architecture. The IAN model was proposed for interactive learning of attentions in the context vector and special tokens (i.e. hate targets in our study), and generate the representations for the special tokens and contexts separately. The IAN model has shown high performance results in many tasks such as aspect-level sentiment classification [28], adverse drug reactions [1], pedestrian detection [42], and other classification tasks.

Unlike the existing methods for hate speech detection task that mainly work for monolingual data or require special linguistic resources created with labor-intensive efforts, we propose to adapt this IAN model. It can facilitate an approach for multilingual hate speech detection with integration of the human-in-the-loop paradigm to improve both the model performance and interpretability. The role of human agents is to provide more contextual information

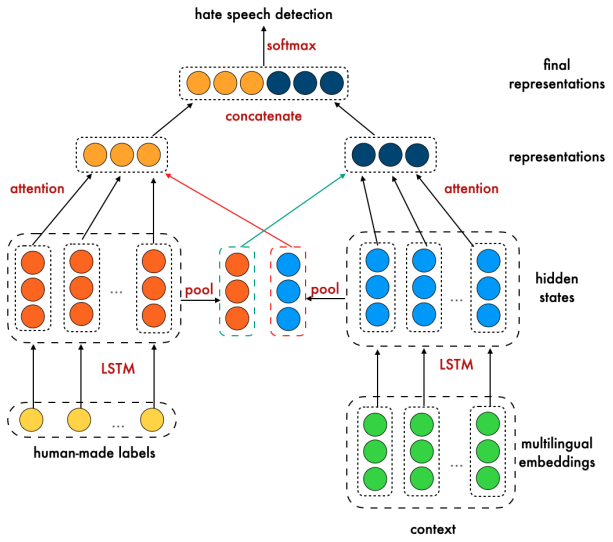

Figure 2: The overall architecture of MLIAN model.

about hate speech that could be informed by appropriately invoking the correct frame. Specifically, we provide whether a given text message has personal or group hate target as an additional parameter for interactive model training. We choose this feature because it can be extracted in real cases creatively – users of the social media platforms can provide the appropriate frame of interpretation and identify the hate target to themselves and label the posts at scale. We evaluate our proposed model against state-of-the-art baseline methods in Section 4.

3.2 Model Architecture

The overall architecture of MLIAN model is shown in Figure 2. MLIAN model contains two parts that interactively model the targets (left part in the Figure 2) and context (right part in the Figure 2).

Our specific steps are as follows. First, we use multilingual text embeddings as input to LSTM. It is employed to obtain a target and hidden states of words on the word level for targets and context respectively. Second, we calculate the average value of the targets' states and the contexts' *hidden states* to supervise the generation of attention vectors, with which the attention mechanism is adopted to capture the important information in the context by the target provided by human feedback. This type of architecture design [28] enables to capture the influence on the context from the identified target and the influence on the target from the context. This approach provides more clues to the modeling algorithm to pay attention to the contextually-relevant hate speech features and thus, allows to generate their effective data representations interactively. Finally, target and context representations are concatenated as a *final representation* for an input text that is fed to a softmax function for hate speech classification.

3.3 Transformer-Based Multilingual Embeddings

We use embeddings generated by two pre-trained transformer-based models for representing the input data for MLIAN: Language-Agnostic SEntence Representations (LASER) [4] and Distilled Multilingual Bidirectional Encoder Representations from Transformers (DistilmBERT). The main difference between these two models is that LASER generates sentence-level embeddings while DistilmBERT generates word/token-level embeddings.

LASER. Results of previous large-scale analysis of multilingual hate speech detection in 9 languages from 16 different sources demonstrate that simple models such as LASER embeddings with machine learning algorithms perform with the best results [3]. LASER is based on an architecture to learn joint multilingual sentence representations for data in 93 languages. Given an input sentence, LASER provides sentence embeddings which are obtained by applying max-pooling operation over the output of a Bidirectional LSTM (Bi-LSTM) encoder. Bi-LSTM output is constructed by concatenating outputs of two individual LSTMs working in opposite directions (forward and backward). This way more contextual information is included in the output than a single LSTM reading text from left to right. The system uses a single Bi-LSTM encoder with a shared byte-pair encoding (BPE) vocabulary for all languages, coupled with an auxiliary decoder, and trained on publicly available parallel corpora. In our experiments, all sentences are initialized by LASER in 1024-dimension fixed-size vector to represent the input textual post. The resulting embeddings are computed using English annotated data only, and transferred to any of the 93 languages without any modification. Experiments in cross-lingual natural language inference (XNLI dataset), cross-lingual document classification (MLDoc dataset), and parallel corpus mining (BUCC dataset) have shown the effectiveness of the LASER approach [4].

DistilBERT. Multilingual DistilBERT model pre-trained by HuggingFace[1] is a distilled version of the multilingual BERT-base model [38]. The model is trained on the concatenation of Wikipedia in 104 different languages. The model has 6 layers, 768 dimension and 12 heads, totalizing 134M parameters (compared to 177M parameters for the multilingual BERT). On average DistilmBERT is 60% faster than multilingual BERT model. All DistilmBERT embeddings have 512-dimension fixed-size vector representation. BERT's key technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modeling [13]. This is in contrast to previous efforts which looked at a text sequence either from left-to-right or combined left-to-right and right-to-left training. We use BERT because it shows high performance in many NLP-tasks including hate speech classification [15, 34, 37].

3.4 Human Feedback Guided by Frame Semantics Theory

While human could give a variety of feedback to the MLIAN model for hate speech detection, we propose to guide the nature of the feedback. One common approach of human-in-the-loop machine learning paradigm is to seek feedback from humans as the correct labels at the entire text level. Alternative to this approach can be

[1] https://huggingface.co/models

the feedback at the level of word tokens in the text, but the question is how to create a principled approach to identify the special tokens for the feedback. We explore frame semantics theory from cognitive linguistics [17] that can help better explain and guide the types of special tokens to focus for the human feedback, in order to rectify the model reasoning provided through attention weight map such as shown in Figure 1. According to the theory of frame semantics, word meanings are defined relative to frames in a given text. Given the varied ways in which hate speeches are expressed, human, rather than machine, could quickly identify the appropriate frame to interpret a given text. Thus, if the model can learn to give attention to the elements of the correct frame as per human interpretation, the model performance could improve.

A semantic frame is a set of statements that give "characteristic features, attributes, and functions of a denotatum (data object), and its characteristic interactions with things necessarily or typically associated with it." [2] Moreover, a semantic frame can be viewed as a coherent group of concepts such that complete knowledge of one of them requires knowledge of all of them [36]. Therefore, it provides a common representation to capture both knowledge and meaning of a given textual post. For example, a description of frame in FrameNet[2] (the popular knowledge base to understand human language) primarily contains following attributes: *Description* - a textual description of the frame including what it represents; *Frame Elements (FE)* - additional attributes for representing meaning of the frame in a sentence/context, such as the frame *Being_born* has FEs: Child, Time, Place, etc.; *Lexical Units (LU)*: the lemmatized form of words with their part-of-speech that invoke a frame; and lastly, *Example Sentences*.

Hate speech can be described by several frames [14], and thus, a common but essential pattern to guide the human feedback at the token-level could be the element of hate target group in a given text. To achieve this goal we propose a model described in the next subsection that can combine human feedback of the hate target interactively during the training process, within the specific language context of a textual post.

4 EXPERIMENT SETUP

Data: We evaluated our model on the dataset containing English and Spanish tweets provided by SemEval-2019 Task 5 — HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter [7]. This dataset has been labeled with two classes for determining whether a given tweet is hateful or not-hateful for a given target such as women or immigrants. Additionally, the data includes labels for identifying the harassed target as individual or generic (i.e. individual or group). Table 2 shows the quantity of training and test instances for each category. In the current work, we named the first type of labeling as hate speech labels, and the second type as target-type labels.

Schemes: To compare our proposed method, given there is not exactly comparable prior work for our multilingual problem setup tested on the same dataset, we construct multiple baseline schemes using classical machine learning models [35] that use LASER embeddings as input features. We didn't use state-of-the-art models proposed during SemEval-2019, because participants had another task and tested their approaches on single-language data, while our setup includes only multilingual cases. Additionally, we also compare MLIAN with LSTM [30] based model, as used in the prior works (we use it with two hidden levels and trained for 20 epochs, similar to MLIAN). We evaluate our MLIAN model with both LASER and DistilmBERT embeddings. The full list of proposed modeling schemes for evaluation is the following (* denotes our proposed models and others are the baselines):

- **[*SVC+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a Linear Support Vector Classifier model.
- **[*RF+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a Random Forest model.
- **[*SGD+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a Stochastic Gradient Descent model.
- **[*MLP+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a Multi-Layer Perceptron model.
- **[*LSTM+LASER*]**: This method uses pre-trained LASER embeddings, which are passed as input to a Long Short-Term Memory Network model.
- **[*LSTM+DistilmBERT*]**: This method uses pre-trained DistilmBERT embeddings, which are passed as input to a Long Short Term Memory Network model.
- **[**MLIAN+LASER*]**: Multilingual Interactive Attention Network method with LASER embeddings.
- **[**MLIAN+DistilmBERT*]**: Multilingual Interactive Attention Network method with DistilmBERT embeddings.

To evaluate the performance of classification models, we adopt three metrics: Accuracy (ACC), Area Under the Receiver Operating Characteristic Curve (AUC), and weighted F-measure (F1), which is consistent with the prior works on hate speech detection.

Model Implementation: In MLIAN, we need to optimize all the parameters in LSTM networks: the attention layers, the softmax layer, and the text embeddings (LASER or DistilmBERT). Cross entropy with L2 regularization is used as the loss function. We use backpropagation to compute the gradients and update all the parameters of LSTM. The coefficient of L2 normalization in the objective function is set to 10^{-3}, the dropout rate is set to 0.2, and 20 epochs.

Table 2: Train and test data.

		Hate speech		Target type	
		Hate	Non-hateful	Individual	Group
Train	English	3783	5217	1341	2442
	Spanish	1857	2643	1129	728
Test	English	427	573	219	208
	Spanish	222	278	137	85

5 RESULT ANALYSIS AND DISCUSSION

We first discuss the results of MLIAN model against the baseline schemes, followed by an in-depth analysis of the nature of human

[2]https://framenet.icsi.berkeley.edu/fndrupal/

feedback, the amount of human feedback, and the analysis of cross-lingual and cross-target scenarios.

5.1 MLIAN Performance

Table 3 shows the performance comparison of *MLIAN*-based models with other baselines. We can observe that the deep learning models have higher performance in multilingual hate speech detection than classical machine learning based model schemes. Further, both the proposed models of *MLIAN+DistilmBERT* and *MLIAN + LASER* show higher performance than LSTM baselines in all metrics. *MLIAN+LASER* model scheme demonstrates better performance result in multilingual cases, which is perhaps contributed by the consideration of sentence-level context by the LASER embeddings and the human feedback in MLIAN.

Table 3: Comparison with baselines. Results of binary classification for the SemEval-2019 Task 5 (hate speech against immigrants and women). Best performances are in bold. Models were trained on multilingual data (10-fold CV). ∗ denotes the proposed models.

Model Scheme	ACC	AUC	F1
MLP+LASER	60.93±0.72	58.73±0.69	59.85±0.64
RF+LASER	70.20±0.61	67.63±0.53	68.85±0.62
SGD+LASER	69.38±0.43	70.07±0.99	70.07±0.52
SVC+LASER	71.87±0.71	71.27±0.45	71.85±0.63
LSTM+DistilmBERT	73.85±0.31	78.29±0.14	73.86±0.31
LSTM+LASER	71.59±0.43	79.38±0.11	71.58±0.43
**MLIAN+DistilmBERT*	81.24±0.59	79.84±0.61	81.00±0.55
**MLIAN+LASER*	**85.06±0.40**	**84.14±0.54**	**84.94±0.42**

Further, our *MLIAN* models demonstrate that emphasizing the importance of human feedback through learning target and context representation interactively can be valuable for hate speech detection tasks. Compared with *LSTM* models, our architecture improves AUC performance by about 6% for LASER-based model implementation and 2% for DistilmBERT-based model implementation on multilingual data. The main reason for higher performance is that *MLIAN+DistilmBERT* and *MLIAN+LASER* use the additional feature contributed by the simulated human agent which can influence the learning of context in the attention network. Besides higher performance, in this design, we can learn the representations of targets and contexts whose collocation contributes to hate speech detection even in the posts where users resort to special language subtleties. This also inspires our future work to explore and research the semantics of a variety of target types.

5.2 Analysis of Frame Semantics Theory-based Human Feedback

In this analysis, we tested the theoretical justification of the impact of human feedback based on the connection between frame semantics theory described in Section 3.4 and attention weight maps resulting from the developed MLIAN model.

Specifically, to understand how MLIAN attention weights correlate with semantic frames, we examine the tweet text originally written in Spanish and its English translation. First, we can extract semantic frames for the text versions in both languages by employing a semantic parser. Semantic parsers are trained specifically to consider context when identifying frames in a text. In this exercise, we utilized SLING [36] to extract head frames from a textual post. Head frames are the frames directly evoked by a mention in text. Figure 3 illustrates the high-level frame extraction process with an example.

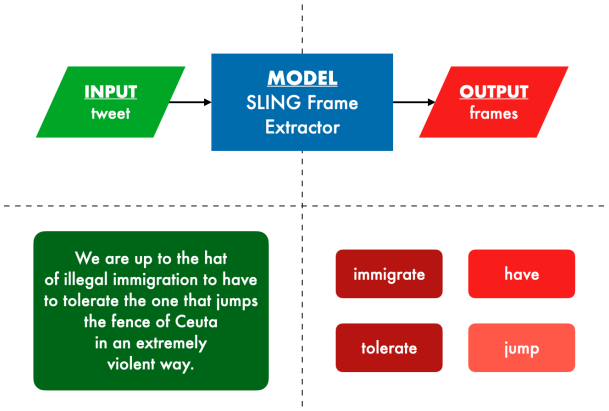

Figure 3: High Level Architecture of Frame Extraction Process.

Second, we apply the trained MLIAN model with and without human feedback data on the two versions of the text to retrieve the attention weight maps. For this task, we used *MLIAN+DistilmBERT* model because it allows to extract word embeddings (while *MLIAN+LASER* extract only sentence embeddings that cannot be interpretable by human.) Figure 4 demonstrates that the attention weights for the MLIAN model with human feedback-based hate targets have more correlations with frames than the model without such principled feedback. Moreover, it is important to note that this correlation is observed for both languages, indicating the significance of relying on a principled approach of frame semantics theory to identify the type of targets to receive the human feedback.

5.3 Analysis of the Impact of Human Feedback

To measure the minimal required human feedback that could impact the MLIAN model performance, we start with a baseline model scheme without considering the target-types labels and then, design several model schemes that incrementally add a specific amount of the target-types labels as human feedbacks. Specifically, we analyze the baseline case against three schemes, with the gradual addition of randomly selected target-type labels – *100, 500, 1000*. For this task, we use the performance measures of Accuracy, F1, and AUC for assessing different model schemes.

The full set of results are presented in Table 4. Best performances are in bold. Results show that the statistically significant improvement of accuracy and AUC was noticeable when increasing the number of target-type labels to just 500, which equals to approximately 4% of training data only. This analysis demonstrates how even the minimal use of human feedback could result into faster convergence of the model training, for better performance in the hate speech detection task.

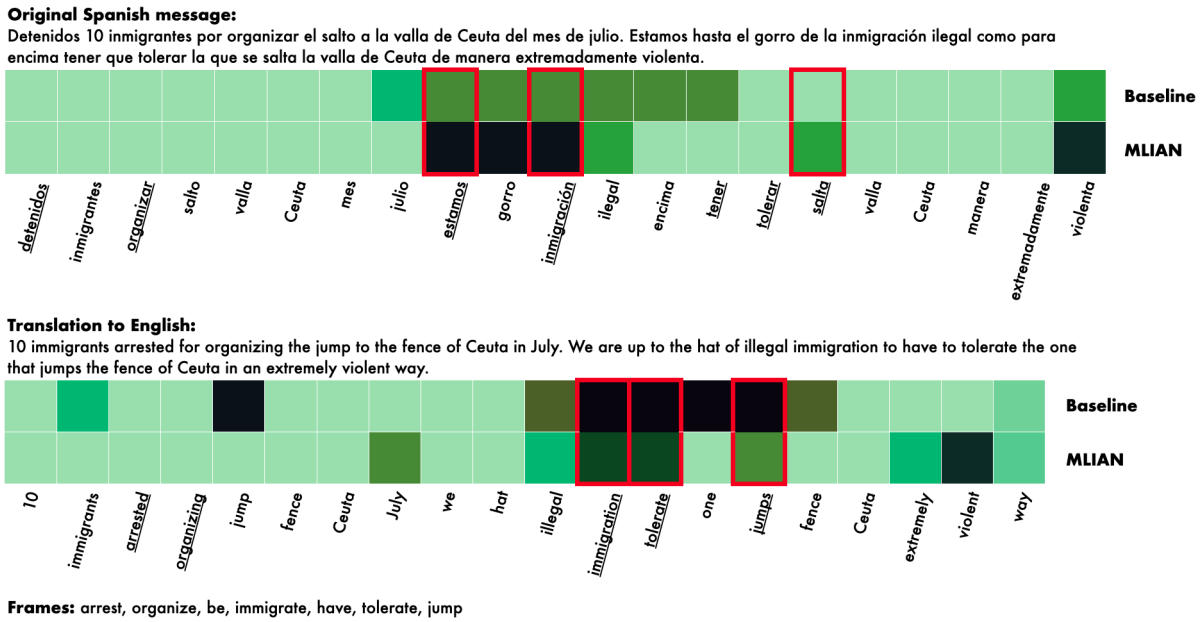

Figure 4: Attention weight maps of the original and translated texts in English and Spanish. The darker color of the word means the higher weight. Cells with red borders indicates words that match with semantic frames.

Table 4: Minimal human impact in *MLIAN+LASER* model is studied for the model schemes with an increasing number of human feedbacks. The table shows *p*-value for each scheme's performance comparison with the baseline and the previous scheme. The best performing scheme's *p*-value is bold.

	baseline	100 feedbacks	500 feedbacks	1000 feedbacks
ACC	73.04	74.10	77.46	78.59
Compare:				
- baseline		0.297	0.001	0.000
- scheme		0.297	**0.028**	0.387
AUC	73.03	74.26	77.02	78.47
Compare:				
- baseline		0.206	0.002	0.000
- scheme		0.206	**0.060**	0.296
F1	73.15	74.16	77.44	78.62
Compare:				
- baseline		0.318	0.001	0.000
- scheme		0.318	**0.031**	0.368

5.4 Analysis of Cross-Lingual and Cross-Target Classification

Multilingual classification tasks also include cross-lingual classification settings — when languages in training and testing data are different.

For evaluation of cross-lingual capability of the proposed method, we train MLIAN model on one language and test on another, for English and Spanish. The full results of cross-lingual classification are presented in Table 5. The MLIAN models for both languages show better results comparing with LSTM baselines. The model achieves up to 68% AUC for training on English and testing on Spanish, while for the opposite scenario, it reaches up to 78% AUC. On the other hand, the LSTM-based baseline models show only 65% and 68% AUC for the two scenarios.

Table 5: Results of cross-lingual classification. Best performances are in bold (10 fold CV). ∗ denotes the proposed models.

Model Scheme	ACC	AUC	F1
EN → ES			
LSTM+LASER	61.90	65.67	61.74
LSTM+DistilmBERT	59.16	65.01	58.77
MLIAN+DistilmBERT*	71.33	**68.16**	69.78
MLIAN+LASER*	**71.46**	**68.16**	**69.91**
ES → EN			
LSTM+LASER	58.16	63.51	57.61
LSTM+DistilmBERT	63.45	68.56	63.13
MLIAN+DistilmBERT*	**81.76**	**78.95**	**81.01**
MLIAN+LASER*	81.28	78.57	80.57

Lastly, we evaluate MLIAN for cross-topic hate speech detection — when the type of targets in the training and testing data are different. For evaluation of cross-target capability of the model, we train the model on the dataset with hate speech targeted to one group of people (e.g. migrants) and tested on the dataset with another targeted group (e.g. women). Example of hate speech targeted at

different groups are presented in Table 1. The full results of cross-topic classification is presented in Table 6. MLIAN models show the best results comparing with another baseline model schemes. MLIAN model reaches up to 80% AUC for training on migrants-as-target dataset and testing on women-as-target dataset, and in the opposite scenario, such a model reaches up to 82% AUC. In contrast, LSTM baselines were only able to achieve up to 74% AUC in both scenarios.

These results show that MLIAN could reach good performance during both cross-lingual and cross-topic hate speech classification tasks and this analysis validates the benefits of deep learning model with human feedback for improving the task performance.

Table 6: Results of cross-target classification. Best performances are in bold (10 fold CV). ∗ denotes the proposed models.

Model Scheme	ACC	AUC	F1
migrants → women			
LSTM+LASER	68.16	74.35	68.16
LSTM+DistilmBERT	68.31	74.46	68.31
MLIAN+DistilmBERT*	71.41	69.37	71.05
MLIAN+LASER*	**81.89**	**80.31**	**81.51**
women → migrants			
LSTM+LASER	67.17	73.33	67.17
LSTM+DistilmBERT	67.23	74.12	67.23
MLIAN+DistilmBERT*	**84.06**	**82.70**	**83.80**
MLIAN+LASER*	70.40	77.25	69.39

6 CONCLUSION AND FUTURE WORK

In this paper, we design a multilingual interactive attention network (MLIAN) model for hate speech detection in social media posts, regardless of language. The core idea of MLIAN is to use two attention networks to model the context of content and the special tokens as targets interactively, where we employ the frame semantics theory to design a principled approach for appropriately guiding the human feedback to provide target labels. We use simulated human feedback by labeling posts that contain personal/group hate for identifying the special tokens as target labels. The model pays close attention to such important parts in the context and learns to give higher attention to the potential elements of the semantic frame characterizing the hate speech in the post. Experiments on SemEval-2019 Task 5 dataset demonstrate that MLIAN model performs better than several baselines and requires a minimal human feedback effort for improving the model performance. We present extensive analyses to show the value of modeling with human feedback, which can help adapt the model to different languages and tasks easily. The application of MLIAN model can inform future studies for multilingual hate speech analytics.

We acknowledge certain limitations of this study that could be addressed in future work. First, the dataset for experimentation contains hate speech directed to two different groups of people, which could be extended to different types of groups and it would be valuable to understand how human feedback for the variety of group targets could affect the model performance. Second, our experiments are based on only English and Spanish language posts due to the dataset limitation, which could be further expanded for the datasets of more languages. Third, the proposed method was tested using a simulated environment of human feedback as we planned to conduct several analyses presented in this paper, but our approach could be tested easily with real human agents in the future as well.

Reproducibility: Datasets and code for the experiments described in this paper will be available for research purposes at public repository https://github.com/vitiugin/mlian.

Acknowledgments: Purohit acknowledges the U.S. National Science Foundation Award IIS-1657379 for partial research support. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

REFERENCES

[1] Ilseyar Alimova and Valery Solovyev. 2018. Interactive attention network for adverse drug reaction classification. In *Conference on Artificial Intelligence and Natural Language*. Springer, 185–196.
[2] Keith Allan. 2001. Natural language semantics. (2001).
[3] Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep Learning Models for Multilingual Hate Speech Detection. *arXiv preprint arXiv:2004.06465* (2020).
[4] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* 7 (2019), 597–610.
[5] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 759–760.
[6] Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Aggression Identification in English, Hindi and Bangla Text using BERT, RoBERTa and SVM. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA), Marseille, France, 76–82. https://www.aclweb.org/anthology/2020.trac-1.12
[7] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 54–63.
[8] Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science* 5, 1 (2016), 11.
[9] Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)* 20, 2 (2020), 1–22.
[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.
[11] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 11–20.
[12] Tom De Smedt. 2020. Profanity & Offensive Words (POW). (2020).
[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*. 4171–4186.
[14] Mark R Dixon, Simon Dymond, Ruth Anne Rehfeldt, Bryan Roche, and Kimberly R Zlomke. 2003. Terrorism and relational frame theory. *Behavior and Social Issues* 12, 2 (2003), 129–147.
[15] Suman Dowlagar and Radhika Mamidi. 2021. HASOCOne@ FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection. *arXiv preprint arXiv:2101.09007* (2021).
[16] Facebook. [n.d.]. *Community Standards. Objectionable Content.* https://www.facebook.com/communitystandards/objectionable_content
[17] Charles J Fillmore et al. 2006. Frame semantics. *Cognitive linguistics: Basic readings* 34 (2006), 373–400.

[18] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30.

[19] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*. 105–114.

[20] Andrea Galassi, Marco Lippi, and Paolo Torroni. [n.d.]. Attention in Natural Language Processing. ([n. d.]).

[21] Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. arXiv:1710.07395 [cs.CL]

[22] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All You Need is" Love" Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. 2–12.

[23] Radhouane Guermazi, Mohamed Hammami, and Abdelmajid Ben Hamadou. 2007. Using a semi-automatic keyword dictionary for improving violent Web site filtering. In *2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*. IEEE, 337–344.

[24] Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael Paul. 2020. Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 1440–1448.

[25] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[26] Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, and Maheshwar Reddy Chennuru. 2018. Trac-1 shared task on aggression identification: Iit (ism)@ coling'18. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. 58–65.

[27] Leonard Williams Levy, Leonard W Levy, Kenneth L Karst, and Adam Winkler. 2000. Encyclopedia of the american constitution. (2000).

[28] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 4068–4074.

[29] Aditya Mahajan, Divyank Shah, and Gibraan Jafar. 2020. *Explainable AI approach towards Toxic Comment Classification*. Technical Report. EasyChair.

[30] Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. 467–472.

[31] Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence* 30, 2 (2018), 187–202.

[32] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*. 14–17.

[33] Ricardo Martins, Marco Gomes, José João Almeida, Paulo Novais, and Pedro Henriques. 2018. Hate speech classification in social media using emotional analysis. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE, 61–66.

[34] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one* 15, 8 (2020), e0237861.

[35] Oluwafemi Oriola and Eduan Kotzé. 2020. Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. *IEEE Access* 8 (2020), 21496–21509.

[36] Michael Ringgaard, Rahul Gupta, and Fernando CN Pereira. 2017. SLING: A framework for frame semantic parsing. *arXiv preprint arXiv:1710.07032* (2017).

[37] Niloofar Safi Samghabadi, Parth Patwa, PYKL Srinivas, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. 126–131.

[38] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[39] Yasas Senarath and Hemant Purohit. 2020. Evaluating Semantic Feature Representations to Efficiently Detect Hate Intent on Social Media. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*. IEEE, 199–202.

[40] André Calero Valdez and Martina Ziefle. 2018. Human factors in the age of algorithms. understanding the human-in-the-loop using agent-based modeling. In *International Conference on Social Computing and Social Media*. Springer, 357–371.

[41] Cindy Wang. 2018. Interpreting neural network hate speech classifiers. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 86–92.

[42] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. 2019. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion* 50 (2019), 20–29.