Towards a Hate Speech Index with Attention-based LSTMs and XLM-RoBERTa

Mauro Bruno[1,†], Elena Catanese[1,†] and Francesco Ortame[1,*,†]

[1] Italian National Institute of Statistics (Istat)

Abstract

The diffusion of hate speech on social media requires robust detection mechanisms to measure its harmful impact. However, detecting hate speech, particularly in the complex linguistic environments of social media, presents significant challenges due to slang, sarcasm, and neologisms. State-of-the-art methods like Large Language Models (LLMs) demonstrate strong contextual understanding, but they often require prohibitive computational resources. To address this, we propose two solutions: (1) a bidirectional long short-term memory network with an attention mechanism (AT-BiLSTM) to enhance the model's interpretability and natural language understanding, and (2) fine-tuned multilingual robustly optimized BERT (XLM-RoBERTa) models.

Building on the promising results from EVALITA campaigns in hate speech detection, we develop robust classifiers to analyse 20.4 million Tweets related to migrants and ethnic minorities. Further, we utilise an additional custom labeled dataset (IstatHate) for benchmarking and training and we show how its inclusion can improve classification performance. Our best model outperforms top entries from previous EVALITA campaigns. Finally, we introduce Hate Speech Indices (HSI), which capture the dynamics of hate speech over time, and assess whether their main peaks correlate with major events.

Keywords

hate speech detection, deep learning, attention mechanism, RoBERTa, artificial intelligence

1. Introduction

Social media platforms provide a fertile ground for the dissemination of hate speech, particularly targeting vulnerable groups such as migrants and ethnic minorities. In the last decade, hateful speech on platforms like X has become a pressing issue, as it not only affects the individuals who are directly targeted, but also contributes to a climate of hostility and division. Detecting hate speech in social media content is crucial to analyse the safety and inclusivity of online platforms and social environments.

Hate speech detection is inherently challenging due to the subtle and evolving nature of social media language. Tweets often contain slang, neologisms, and sarcasm, which complicates the identification process. Traditional text classification methods usually fall short in addressing these challenges, especially for non-English languages where extensively labeled training sets are not easy to gather, calling for the development of more sophisticated approaches.

The topic of hate speech detection in Italian texts has gained significant attention within the natural language processing (NLP) community, as shown by the *HaSpeeDe* (Hate Speech Detection) tasks at EVALITA. For instance, the EVALITA 2018 [1], and 2020 [2] campaigns have provided labeled datasets and attracted several submissions employing a diverse set of machine learning and deep learning techniques. A prominent approach in recent hate speech detection and, in general, text classification, is the use of pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT) [3]. After their first appearance in 2018, BERT-based models have set new standards in several NLP tasks thanks to their ability to capture contextual information effectively, especially when fine-tuned on the specific task of interest. In 2019, a multilingual robustly optimized BERT (XLM-RoBERTa) [4] was published, making it possible to obtain higher performances on non-English texts. For instance, *TheNorth* team for the *HaSpeeDe 2* task at EVALITA 2020 obtained the best results fine-tuning a XLM-RoBERTa model [5].

It is also worth noting that in recent years, generative Large Language Models (LLMs) have demonstrated an even more impressive ability to understand natural language. However, their large number of parameters makes them impractical for classifying large volumes of data, even when compared to the larger version of XLM-RoBERTa[1]. Given these developments and challenges, our research proposes two approaches to hate speech classification: (1) an attention-based bidirectional long short-term memory network (AT-BiLSTM), benchmarked against a standard BiLSTM model, and (2) a fine-tuned

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

†These authors contributed equally.

mbruno@istat.it (M. Bruno); catanese@istat.it (E. Catanese); francesco.ortame@istat.it (F. Ortame)

0009-0007-8672-872X (F. Ortame)

[1] The number of parameters in Large Language Models ranges between a few billions to hundreds of billions of parameters, while the large version of XLM-RoBERTa "only" has 561 million parameters.

XLM-RoBERTa (large) model, benchmarked against its base, smaller version. We use two labeled training sets: (a) the EVALITA 2020 *HaSpeeDe 2* task dataset, and (b) a custom, smaller labeled dataset, which we refer to as *IstatHate*. Our study explores the impact of training models on both the EVALITA dataset alone and a combined dataset that includes EVALITA and *IstatHate*, evaluating their performance across multiple test sets.

Finally, we present a preliminary version of the Hate Speech Index (HSI), designed to quantify the proportion of hate speech by classifying 20.4 million Italian Tweets related to migrants and ethnic minorities from January 2018 to February 2023.

2. Data

This section describes the data used for training, validating, and testing the models and the corpus of Tweets on which we compute the hate speech index (HSI).

2.1. Corpus

The prediction corpus consists of 20.4 million unlabeled Tweets from January 2018 to February 2023. The Tweets are obtained through a two-step filtering procedure: *first*, a general 250-keyword filter gathers Tweets directly from X's API;*second*, a smaller, immigration-related keyword filter retrieves the relevant Tweets from the database. Thematic experts, borrowing the contents of discrimination survey questionnaires, have derived a preliminary filter. These regular or stemmed expressions have been validated by means of topic modelling analysis and word embedding. For instance, the word *cinese* ("chinese") was almost always related to markets or products and has therefore been removed. We also noticed that due to the generic term *stranieri* ("foreigners") there are also some residual out-of-scope and irrelevant conversations. These issues only affects around 5% of the total texts. The final filter consists in 21 stemmed expression (ex. *immigrat-*), or complete words.

2.2. Training data

EVALITA Most of the labeled training data comes from the EVALITA 2020 *HaSpeeDe 2* task. The distribution of the labels in the training dataset is shown in Table 1.

IstatHate Additionally, we use a custom-labeled dataset, i.e., *IstatHate*, derived from our corpus in the following way: (1) we fit a Latent Dirichlet Allocation (LDA) model [6] on the entire corpus, (2) we identify clusters likely to contain hateful Tweets, i.e., those with offensive language, such as "*fate schifo*" ("you suck"), and "*avete rotto i c****oni*" ("you pi**ed us off") and few others, (3) we retrieve Tweets from these clusters, identifying the expressions with a probability of 1 of belonging to the clusters. This approach isolates 242,000 Tweets, of which 67,000 are unique. It is worth noticing that viral Tweets (the ones that are repeated/retweeted several times) need to be annotated with a higher probability. A common practice to draw a much more efficient sample instead of simple random sampling is to use stratified sampling, an effective method for handling skewed distributions. In particular, we adopted [7]. (4) We employ stratified sampling using the total number of Tweets as the target variable, and we divided that variable into five classes using them as stratification criteria. (5) The Tweets are then stratified into the classes based on the number of retweets, with the final class being a take-all stratum, resulting in 681 sampled texts, ensuring a coefficient of variation of 5%. (6) These 681 Tweets are then manually labeled by Istat researchers adopting the following criteria: if the language is vulgar/aggressive but generic it is not labeled as hateful, if, on the contrary, it is related to migrants and/or ethnic minorities and the hate/prejudice is clearly directed towards them, then they were labeled as hateful. The weighted estimate indicates that 34% of the Tweets contains hateful language, serving as a rough upper bound of the hate proportion within our prediction corpus. Even if our sample dataset likely over-represents hateful content, we disregard the weighting at this preliminary phase, simply adding *IstatHate* to the EVALITA dataset.

Table 1
Labeled data distribution

dataset	split	n	% hateful	% not hateful
EVALITA	train	5469	40,46%	59,54%
	eval	1368	40,42%	59,58%
	test	1263	49,25%	50,75%
IstatHate	train	435	33,79%	66,21%
	eval	137	29,93%	70,07%
	test	109	33,94%	66,06%
Full	train	5904	39,97%	60,03%
	eval	1505	39,47%	60,53%
	test	1372	48,03%	51,97%

Table 1 shows the distribution of the labeled data between *hateful* and *not hateful* Tweets and across datasets and splits.

3. Methodology

In this section, we present the methodology adopted in our study and outline the experimental design. We begin by introducing the model architectures, followed by a detailed description of the training procedure.

3.1. AT-BiLSTM model architecture

The architecture of our attention-based bidirectional LSTM (AT-BiLSTM) model comprises four main components: an embedding layer, a bidirectional LSTM layer, an attention layer, and an output layer. We will detail each component sequentially.

Embedding layer We pre-train a FastText [8] embedding model on the prediction corpus and extract the word vectors to initialise the weights of the embedding matrix. Table 2 presents the main training parameters of our model: each word is represented by a 300-dimensional vector, the training considers a distance window between words of up to 8 positions, and the model is trained for 25 epochs using a continuous bag-of-words algorithm.

Table 2
FastText embedding model hyperparameters.

dim	window	epochs	algorithm
300	8	25	skip-gram

As emerged from the hyperparameter optimization phase[2], we keep the embedding weights fixed during the AT-BiLSTM training.

Attention mechanism In deep learning, attention mechanisms can improve model performance by focusing on important features of input sequences.

In our model, the attention mechanism is implemented on top of the LSTM layer to focus on the most relevant parts of the input sequence for predictions [9]. Our attention mechanism works as follows:

- Transform the LSTM output using a fully connected layer to get attention scores for each word.
- Normalise these scores into attention weights with a *softmax* function, creating a pseudo-probability distribution.
- Compute a context vector by taking a weighted sum of the LSTM outputs using the attention weights. This context vector emphasizes the most important parts of the input sequence for the classification task[3].

The attention mechanism allows our model to dynamically focus on different parts of the input for different examples.

LSTM layer The core of our model is a bidirectional Long Short-Term Memory (LSTM) network. LSTMs are a specialized type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data [10]. The *bidirectional* aspect of our LSTM processes the input sequence in both forward and backward directions. This bidirectionality provides the network with context from both past and future states for any given point (word) in the sequence (sentence) [11]. In practice, this means that when our model is processing a word in a Tweet, it has information about the words that came before and after it, allowing for an increased understanding of context.

The LSTM layer consists of multiple stacked bidirectional LSTM cells. Each cell maintains a cell state and a hidden state, which are updated at each time step as the input sequence is processed. The number of layers is included in the hyperparameter optimization phase.

Output layer The final component of our model is a fully connected (dense) layer that takes the context vector produced by the attention mechanism as input. The output dimension of this layer is one-dimensional, as there are two classes in our hate speech detection class. The output of this layer is passed through a softmax function to produce a number between 0 and 1. Finally, the class is assigned comparing the output with a threshold (0.5).

The optimal configuration for each LSTM-based model, resulting from Bayesian hyperparameter optimization, is detailed in the Appendix.

3.2. XLM-RoBERTa

Multilingual RoBERTa (XLM-RoBERTa, or XLM-R) is a transformer-based model that builds upon the original BERT model and the monolingual RoBERTa (Robustly Optimized BERT Pretraining Approach) model [12]. It is designed to handle multiple languages, making it particularly suitable for our task of hate speech detection in Italian texts.
XLM-RoBERTa is trained on 100 different languages and has a much larger vocabulary size (250k tokens) compared to both BERT (30k tokens) and RoBERTa (50k tokens).

3.3. Training

In this section, we outline the experimental design we followed to obtain our results. We structured our experiments to systematically assess model performance under different training conditions and across various test sets.

[2] We ran both random search and Bayesian optimization. The best result came from the latter.

[3] We also experimented with attention masking. However, this negatively impacted accuracy. Upon inspecting the attention scores, we observed that the model naturally assigns negligible weights to padding tokens.

3.3.1. Experimental design

Training sets We trained each model under *two* distinct scenarios: (1) a training set comprising only data from the EVALITA labeled dataset, and (2) a training set comprising both EVALITA data and *IstatHate* data.

Evaluation We evaluate every model on *three* test datasets: (a) a test set comprising only data from the EVALITA test dataset, (b) a test set comprising only data from the *IstatHate* test set, and (c) a combined test set comprising data from both EVALITA and *IstatHate* test sets. None of the texts in these test sets are seen by the models during training, in any scenario.

Therefore, we have four different architectures and two training sets, resulting in eight distinct models.

3.3.2. Model Training

LSTM-based We ran a Bayesian optimization process to automatically extract optimal hyperparameters. This optimization process is detailed in the Appendix. We trained the models for 10 epochs, and we extracted the best configuration based on validation loss.

XLM-RoBERTa Given the large size of XLM-RoBERTa models, we were not able to run Bayesian optimization, and instead employed grid search over a reduced subset of hyperparameters. We trained the models for 10 epochs, and extracted the weights from the run with the lowest validation loss. We follow a training procedure loosely based on the methodology outlined by [13], but with adaptations to the data and hyperparameters to optimise performance for our specific use case. A detailed description of the training hyperparameters can be found in Appendix A.1.

4. Results

In this section, we present the results of our analysis, covering model performance, attention weight visualizations, and Hate Speech Index (HSI) predictions.

4.1. Model performance

Table 3 highlights the performance of the models, presenting the macro F1 score across the different test sets.

There are several observations that can be made about these results. First, there is a clear positive correlation between model size and performance, particularly evident in the XLM-RoBERTa models, where the larger variant consistently outperforms the smaller ones across all test sets. This is expected for a complex task like hate speech detection.

Table 3
Comparative model performance on different test sets

	Tested on		
Model	Full	EVALITA	IstatHate
BiLSTM-EV	0,761	0,773	0,627
BiLSTM⋆	0,758	0,763	0,690
AT-BiLSTM-EV	0,763	0,780	0,550
AT-BiLSTM⋆	0,773	0,779	0,676
XLM-R-base-EV	0,773	0,788	0,603
XLM-R-base⋆	0,772	0,778	0,672
XLM-R-large-EV	0,796	0,810	0,632
XLM-R-large⋆	**0,811**	**0,816**	**0,750**

$^{(-EV)}$ Trained only on EVALITA
$^{(\star)}$ Trained on both EVALITA and *IstatHate*.

A more interesting observation can be made about the effect of including *IstatHate* in the training set along EVALITA data: besides the expected increased performance on the *IstatHate* test set, there is a case in which the performance on the EVALITA test set increases too, namely XLM-RoBERTa-large⋆. This non-trivial cross-dataset improvement, suggests that training on both datasets enhances the model's generalization capabilities, despite the fact that the datasets were labeled by different people. Finally, it is interesting to notice how a simpler model like AT-BiLSTM⋆ manages to outperform XLM-RoBERTa-base⋆ on all test sets.
Results on the *IstatHate* test set are consistently lower than results on the EVALITA test set, but this was expected, as, even when included in the training, *IstatHate* is much smaller in size.
The *Full* test set is a combination of the EVALITA test set and the *IstatHate* test set, and therefore the macro F1 scores on the *Full* test set are a weighted mean between the ones obtained on EVALITA and *IstatHate*.
The best performing model across all test sets is XLM-RoBERTa-large⋆, i.e. fine-tuned on the training set combining both EVALITA and *IstatHate*.

A detailed table that compares the training and inference times of the different models can be found in Appendix A.2.

4.2. Attention visualization

An advantage of an AT-BiLSTM model over a standard BiLSTM model is its ability to visualise attention scores for each word, making outputs more interpretable[4]. Visualising attention scores provides a useful method for empirically examining the impact of training models on different datasets. For instance, the following are two Tweets classified by the AT-BiLSTM-EV model, along

[4] Attention scores can be visualized in BERT-based models too [14], but the XLM-RoBERTa tokenizer does not always split Italian text into complete words, making interpretation trickier.

with their corresponding attention scores.

Tweet 1 (true: No Hate, predicted: Hate)

IT *poi rompe il caz**o a tutti perché ha accolto una famiglia di profughi*

EN *then they break our ba**s because they hosted a family of refugees*

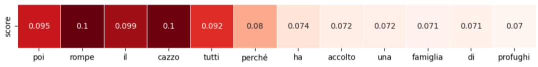

Figure 1: AT-BiLSTM-EV attention scores for Tweet 1.

Tweet 2 (true: Hate, predicted: Hate)

IT *Ipocriti farabutti. Fanno morire i terremotati per i bastardi clandestini immigrati schifosi*

EN *Hypocritical scoundrels. They let the earthquake victims die for the bastard disgusting illegal immigrants.*

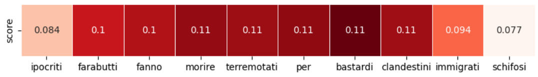

Figure 2: AT-BiLSTM-EV attention scores for Tweet 2.

The first Tweet is misclassified by the AT-BiLSTM-EV model. Analysing the attention scores, we can see how a lot of emphasis was put on curse words both on Tweet 1 and Tweet 2. Figure 3 shows the attention scores produced by the AT-BiLSTM⋆ model for Tweet 1 and Tweet 2, both texts are correctly classified. We can see how a lot of attention is still put on curse words like *ca**o* and *bastardi*, but a significant attention score is also given to *profughi* ("refugees") in Tweet 1. Since the Tweet is correctly classified as *not hateful* – it contains aggressive language but not directed towards migrants or ethnic minorities – we can assume that there is an increased contextual understanding compared to AT-BiLSTM-EV. Additionally, Figure 3 (bottom) shows how the distribution of attention scores for the AT-BiLSTM⋆ model is much more concentrated compared to AT-BiLSTM-EV.

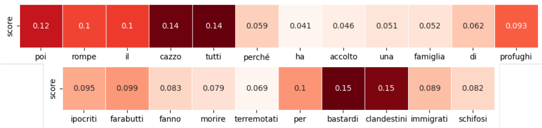

Figure 3: AT-BiLSTM⋆ attention scores for Tweet 1 and 2

4.3. Hate Speech Index (HSI)

In this section, we present and briefly discuss our preliminary Hate Speech Index (HSI) results.

Firstly, the daily HSI is computed as follows:

$$HSI_t = \frac{N_{hate,t}}{N_{hate,t} + N_{nohate,t}},$$

where $N_{hate,t}$ is the number of Tweets classified as hateful on day t, and $N_{nohate,t}$ is the number of Tweets classified as not hateful on day t.

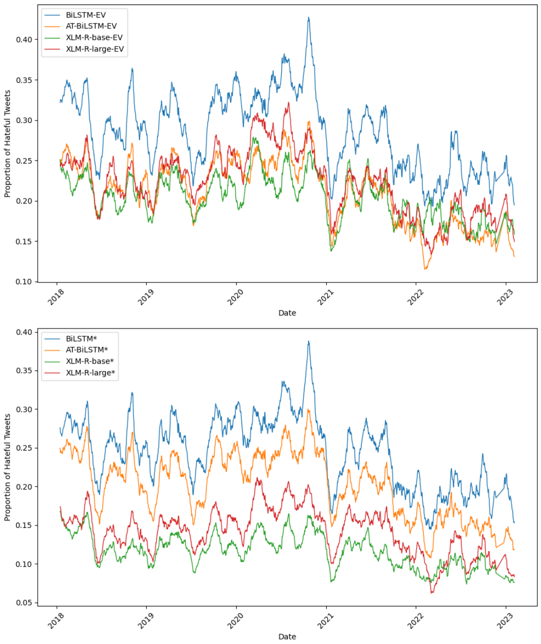

Figure 4: 30-day centered moving average predictions of the models trained only on EVALITA data (top) and on both EVALITA and *IstatHate* (bottom).

Figure 4 displays the different versions of the HSI as derived from the different models.

Descriptive statistics Table 4 illustrates descriptive statistics for the daily HSI.

Table 4
Mean and SD values for HSI.

	EV		⋆	
	mean	sd	mean	sd
BiLSTM	0.285	0.085	0.245	0.081
AT-BiLSTM	0.210	0.071	0.201	0.072
XLM-R-base	0.204	0.063	0.116	0.045
XLM-R-large	0.222	0.071	0.141	0.055

One immediately noticeable difference between the models trained solely on EVALITA and the models trained on

EVALITA and *IstatHate* are the consistently lower levels of the predictions coming from the latter compared to the former for all settings. In particular, the minimum decrease is recorded by BiLSTM models (-0.01), while the maximum decrease is achieved by XLM-RoBERTa-base (-0.09). The lowest mean value for the HSI is achieved by XLM-RoBERTa-base⋆ with an average, indicating a percentage of 11.7% hateful Tweets over the total Tweets in the corpus. The best performing model, XLM-RoBERTa-large⋆, predicts 14.1% of hateful Tweets.

With respect to the standard deviation, we observe that, XLM-RoBERTa models show lower variability compared to LSTM-based models. For XLM-RoBERTa and BiLSTM models, the standard deviation decreases when including *IstatHate* in the dataset.

Correlation The dynamics of the moving averages of the indices appear to be relatively coherent between models, as confirmed by correlations in the range between 0.81 (AT-BiLSTM⋆ vs XLM-RoBERTa-base-EV) and 0.98 (BiLSTM⋆ vs BiLSTM-EV). The lowest correlations between models with the same architecture and different training sets amounts to 0.88 (XLM-RoBERTa-base⋆ vs XLM-RoBERTa-base-EV).

We can now analyse a few peaks in the daily time series to empirically assess the quality of the estimates, and the ability of the models to detect specific events.

October 24, 2018 This date refers to the diffusion of the news about an unfortunate event in which a 16 years old girl was raped and killed by a group of men from Senegal and Nigeria. If we look at the trends in Figure 5 (top) and Figure 6 (top) in Appendix B.1, we notice how the increase in the proportion of hate speech persists in the following period. In this case, we observe that all models detect the event registering values more than twice their average.

July 25, 2021 This peak refers to a news about another 16 years old Italian girl that was beaten up on the street by her 17 years old Moroccan boyfriend. From Figure 5 (bottom) and Figure 6 (bottom) in Appendix B.1, we can see how not all models detect this event. In particular, of the models trained on both EVALITA and *IstatHate*, only XLM-RoBERTa-large⋆ and AT-BiLSTM⋆ show a clear peak in the trend, while LSTM-based models trained only on EVALITA struggle to identify this peak. The only model that detects the peak in both cases is XLM-RoBERTa-large, further empirically confirming its robustness.

We also inspected the negative shift at the beginning of 2021, detected by every model. Analysing the single days it appears that it is more of a trend rather than a response to a specific event/series of events.

5. Conclusion

This study addressed the issue of hate speech detection on social media, specifically focusing on X (formerly Twitter) and on migrants and ethnic minorities. Given the complexities of natural language on these platforms, we explored different approaches including lighter bidirectional LSTM models with and without attention mechanisms, and fine-tuned XLM-RoBERTa models both in their base and large formats. We trained our models on EVALITA 2020 *HaSpeeDe 2* data and also introduced a small labeled dataset, *IstatHate*, that improves the performance of the already best performing model, XLM-RoBERTa-large, when included in the training set.

Despite longer inference times and higher computational resources required for large amounts of data, heavier models like XLM-RoBERTa-large achieve significantly higher performance and generalization capabilities. Yet, AT-BiLSTM⋆ (i.e., the AT-BiLSTM model that includes both EVALITA and *IstatHate* data in the training), outperforms XLM-RoBERTa-base⋆ across all test sets, a notable achievement considering the difference in models size and inference time.

We compared the predictions of AT-BiLSTM-EV against AT-BiLSTM⋆ visualising the attention scores they assigned to the same Tweets. Empirical evidence shows that including *IstatHate* in the training set may improve contextual understanding and mitigate the bias that simpler models like LSTMs may have when classifying hate speech in the presence of curse words.

The preliminary computation of the Hate Speech Index (HSI) reveals significantly different levels of hate speech detection across different models and training sets, even though the training data has very similar characteristics. Fine-tuned XLM-RoBERTa models produce the lower estimates in levels, especially when *IstatHate* is included in the training set. Furthermore, when analysing hate peaks, XLM-RoBERTa-large⋆ predictions highly correlate with major events.

Future work will focus on expanding and validating the *IstatHate* dataset, exploiting the sampling weights, refining model architectures, and exploring additional features to enhance detection capabilities.

References

[1] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, et al., Overview of the evalita 2018 hate speech detection task, in: Ceur workshop proceedings, volume 2263, CEUR, 2018, pp. 1–9.

[2] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2020).

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[4] A. Conneau, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[5] E. Lavergne, R. Saini, G. Kovács, K. Murphy, Thenorth@ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection, Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020) 2765 (2020) 142–147.

[6] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[7] S. Baillargeon, L.-P. Rivest, The construction of stratified designs in r with the package stratification, Survey Methodology 37 (2011) 53–65.

[8] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.

[9] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[10] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[11] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing 45 (1997) 2673–2681.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[13] D. Nozza, F. Bianchi, G. Attanasio, Hate-ita: Hate speech detection in italian social media text, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), 2022, pp. 252–260.

[14] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

A. Optimization

A.1. Hyperparameters

Here, we show the optimal hyperparameters resulting from 50 iterations of Bayesian optimization of 10 epochs each for the LSTM-based models.

Table 5
Optimal hyperparameters for LSTM-based models

model	hid	n	drop	lr	decay	bs
AT-BiLSTM-EV	32	4	0.48	2.7e-3	1.52e-5	32
AT-BiLSTM⋆	128	2	0.40	3.0e-3	2.15e-4	32
BiLSTM-EV	128	3	0.48	1.4e-3	7.23e-6	16
BiLSTM⋆	64	2	0.49	1.1e-3	1.8e-6	32

In Table 5, *hid* represents the hidden dimension of the network, *n* the number of bidirectional LSTM layers, *drop* the dropout rate, *decay* the weight decay and *bs* the training batch size. The entire process took around 15 minutes for each model running on a NVIDIA T4 GPU.

For XLM-RoBERTa models, we used consistent hyperparameters, shown in Table 6.

Table 6
Hyperparameters for XLM-RoBERTa models

model	lr	scheduler	decay	bs	ga-steps
XLM-R	2e-5	linear	0.01	128	2

Where *scheduler* is the learning rate scheduler and *ga-steps* represents the gradient accumulation steps, meaning that instead of updating the weights immediately after each forward and backward pass for every mini-batch, the gradients are kept in memory and accumulated over several (two, in this case) mini-batches, simulating a larger batch size using less memory.

A.2. Training and Inference Time

We detail the training and inference times, grouping the LSTM-based methods in a single category and keeping XLM-RoBERTa (base) and XLM-RoBERTa (large) separated due to the difference in size between the models.

Table 7
Training and inference times

architecture	train	inf	gpu
LSTM	10s	3-8m	T4
XLM-R-base	15m	25m	A100
XLM-R-large	30m	45m	A100

B. Results

B.1. Peaks

Here, we show the daily index of the different models for the dates mentioned in the results section of the paper. The results come from the models trained on both EVALITA and *IstatHate*.

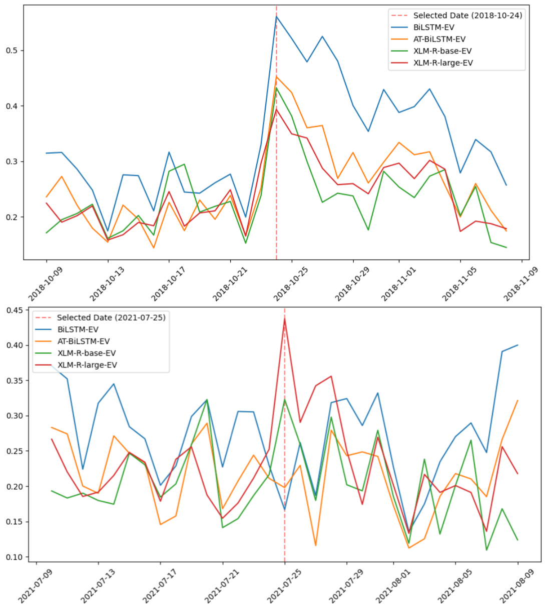

Figure 5: Daily HSI around peaks for models trained only on EVALITA.

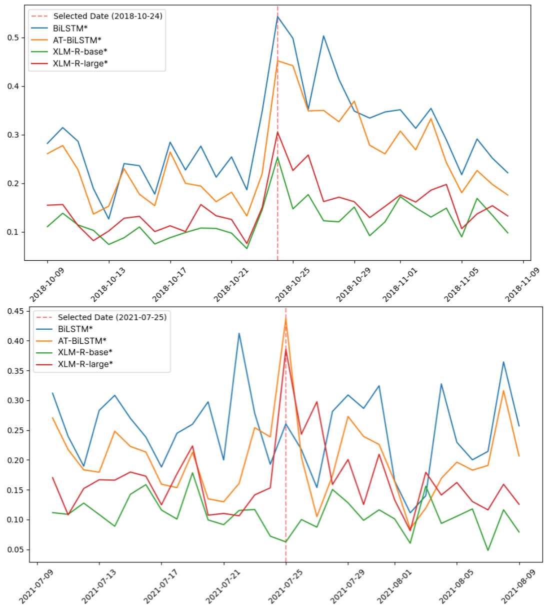

Figure 6: Daily HSI around peaks for models trained on both EVALITA and *IstatHate*.