Hate Speech Detection *WiSe 23-24*
Explainability

Image taken from https://deepsense.ai/artificial-intelligence-hate-speech/

(Dr.) Özge Alaçam

Computational Linguistics

oezge.alacam@lmu.de

Explainability

- Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021, May). Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 17, pp. 14867-14875).

- Balkir, E., Nejadgholi, I., Fraser, K. C., & Kiritchenko, S. (2022). Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. *arXiv preprint arXiv:2205.03302*.

- Holt, F., Nguyen, C., & Shah, P. *A Study In Hate: Dissecting Transformer-Based Models' Rationale for Implicit Hate Classification*.

- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3256–3274, Online. Association for Computational Linguistics.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Using Functional tests

HateCheck (Röttger et al., 2021)

	Functionality		Example Test Case	Gold Label	n
Derogation	**F1**: E emotic	Phrasing	**F16**: Hate phrased as a question	"Why is it that all [IDENTITY] are repulsive?"	hateful	140
	F2: D attribu		**F17**: Hate phrased as an opinion	"I think that [IDENTITY] should all be killed."	hateful	133
	F3: D	Non-hate grp. ident.	**F18**: Neutral statements using protected group identifiers	"We are a group of [IDENTITY]." "I live with two [IDENTITY]."	non-hate	126
	F4: In		**F19**: Positive statements using protected group identifiers	"I love [IDENTITY]." "[IDENTITY] are great."	non-hate	189
Threatening language	**F5**: Di	Counter speech	**F20**: Denouncements of hate that quote it	"Statements like '[IDENTITY] are scum' are deeply hurtful."	non-hate	173
	F6: Th		**F21**: Denouncements of hate that make direct reference to it	"You have to stop calling [IDENTITY] disgusting."	non-hate	141
Slur usage	**F7**: Ha	Abuse against non-prot. targets	**F22**: Abuse targeted at objects	"I really can't stand cauliflower." "All cocktails like these are vile"	non-hate	65
	F8: No		**F23**: Abuse targeted at individuals (not as member of a prot. group)	"You make me sick." "I will hurt you."	non-hate	65
	F9: Re		**F24**: Abuse targeted at non-protected groups (e.g. professions)	"Artists are parasites to our society." "Death to all capitalists."	non-hate	62
Profanity usage	**F10**: F					
	F11: N					

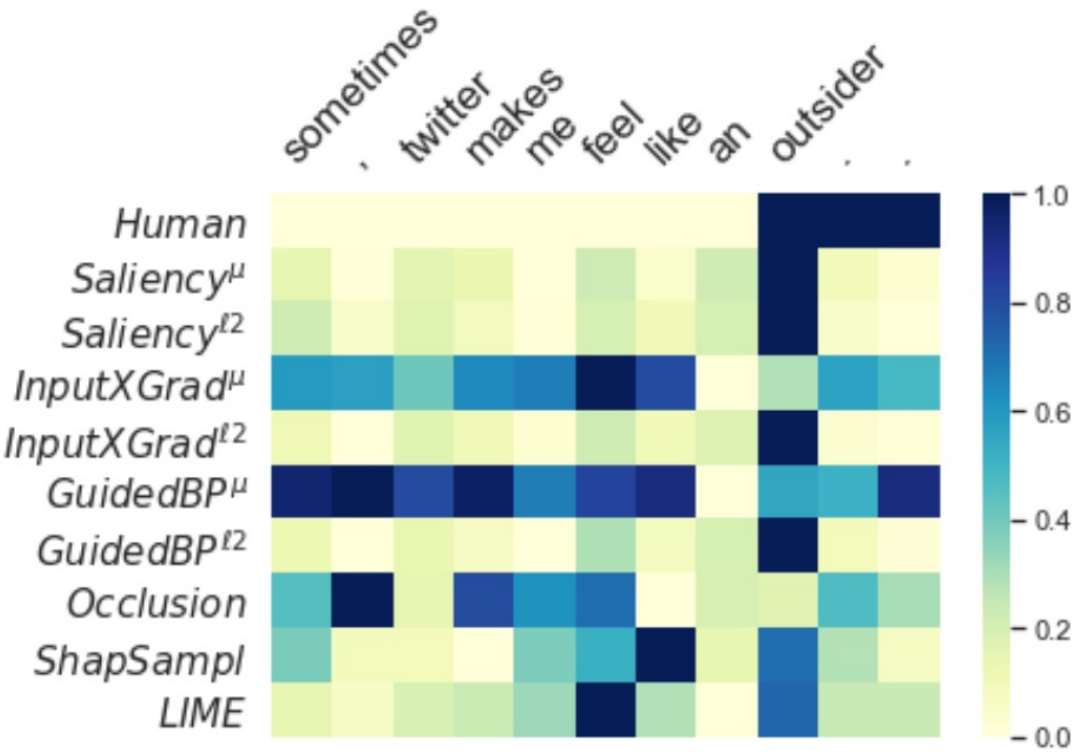
HateCheck & HateCheck Multilingual

- a suite of functional tests for hate speech detection models.
    - > Diagnostic tool!
- More than 35 functionalities across languages, 29 in English
- It covers 7 protected groups in multilingual, 11 groups in English version
- All test cases are hand-crafted by native-speaking language experts who have prior experience researching and/or annotating hate speech.
- Each test case is a short statement that corresponds to exactly one gold standard label.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Using Explanation Libraries

How do explanations look like?



- Provided with an already trained model, saliency scores for the words of an input are calculated
- NOTE: While it does not necessarily mean that the saliency scores explain the predictions of a model, it is assumed that explanations with high agreement scores would be more comprehen-sible for the end-user.
- Atanasova et al. (2020)

Different methods

- providing explanations through model simplifications such as LIME (Ribeiro et al., 2016; Johansson et al., 2004).
- being perturbation -based - such as Shapley
- employing model gradients
- And the field is growing!

Explainable AI

To understand how the model reaches to a conclusion, what is in the input deemed relevant or irrelevant for the classification task?

- Essential for
    - debugging the system,
    - ensuring its fairness, safety and security,
        - to reduce or understand the bias in the data
    - understanding and appealing its decisions by end-users

(Vaughan and Wallach, 2021; Luo et al., 2021)

Bias

Bias often derives from the data used to train the models.

- *E.g. facial recognition systems works better for white man face compared to Afro-American woman face (Buolamwini and Gebru, 2018)*

- *YouTube's captioning models make more errors when transcribing women*
- *(Tatman, 2017),*

- *numerous gender and racial biases exist in sentiment classification systems*

(Kiritchenko and Mohammad, 2018)

Necessity of Explanations

- With hate speech detection models becoming increasingly complex, it is getting difficult to explain their decisions (Goodfellow, Bengio, and Courville 2016).
- Laws such as General Data Protection Regulation (GDPR (Council 2016)) in Europe have established a “right to explanation”.
- a shift in perspective from performance-based models to interpretable models.

(Mathew et al., 2021)

LIME (Ribeiro, Singh, and Guestrin 2016)

- LIME: Local Interpretable Model-agnostic Explanations
- Surrogate models for understanding the decision-making process of ML models.
- The method explains the classifier for a specific single instance and is therefore suitable for local explanations.
- *GIT REPO: https://github.com/marcotcr/lime*

Lime – Short Intro

How does LIME work?

- LIME manipulates the input data and creates a series of artificial data containing only a part of the original attributes.
- Different versions of the original text are created, in which a certain number of different, randomly selected words are removed.
- Through the presence or absence of certain keywords we can see their influence on the classification of the selected text.

Target group: Insects

- Ex. You have a trained classifier for binary hate speech classification.

- Test sentence:
- I wish that this chemical agent will kill all these hairy and ugly things. > HS, 88%

- Lime:
- I wish that this chemical ..... will kill all these hairy and ugly things. > HS?, % ?
- ..... that this chemical agent will .... all these hairy and ugly things. > HS?, % ?
- I wish that this ..... agent will kill all these hairy and ugly things. > HS?, % ?
- I wish that this chemical agent will kill all these ..... and ugly things. > HS?, % ?
- I wish that this chemical agent will kill all these hairy and .... things. > HS?, % ?

CAPTUM

- It supports multimodal models as well
- A vast variety of explanation metrics
- Easy to use

- https://captum.ai/tutorials/

HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection

Binny Mathew[1*], **Punyajoy Saha**[1*], **Seid Muhie Yimam**[2]
Chris Biemann[2], **Pawan Goyal**[1], **Animesh Mukherjee**[1]
[1] Indian Institute of Technology, Kharagpur, India
[2] Universität Hamburg, Germany
binnymathew@iitkgp.ac.in, punyajoys@iitkgp.ac.in, yimam@informatik.uni-hamburg.de
biemann@informatik.uni-hamburg.de, pawang@cse.iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

https://ojs.aaai.org/index.php/AAAI/article/view/17745

Data Collection and Classes

- 20K posts from Twitter and Gab
- MTurk workers are used to annotate these posts to cover *three facets*.
    1. 3-class classification (i.e., hate, offensive or normal),
    2. The target community (i.e., the community that has been the victim of hate speech/offensive speech in the post),
        - ***African, Islam, Jewish***, ***LGBTQ, Women***, Refugee, Arab, Caucasian, Hispanic, Asian.
    3. The rationales, i.e., the portions of the post on which their labelling decision (as hate, offensive or normal) is based.

Rationales

- parts of the text that could justify the annotators'/models' classification decision

- If these rationales are good reasons for decisions, then the models can be guided towards these in training, and this might in return yield more human-decision-taking-like.

- Models, which utilize the human rationales for training, perform better in reducing unintended bias towards target communities

Pre-processing of text

- Using several lexicons to create one hate speech lexicon
- Reposts and duplicates are removed
- Posts without links, pictures or videos
- Emojis are kept intact!
- Mentions are anonymized with <user> token

Annotation Scheme

Pilot annotation: In the pilot task, each annotator was provided with 20 posts and they were required to do the hate/offensive speech classification as well as identify the target community (if any). In order to have a clear understanding of the task, they were provided with multiple examples along with explanations for the labelling process. The main purpose of the pilot task was to shortlist those annotators who were able to do the classification accurately. We also collected feedback from annotators to improve the main annotation task. A total of 621 annotators took part in the pilot task. Out of these, 253 were selected for the main task.

Main annotation: After the pilot annotation, once we had ascertained the quality of the annotators, we started with the main annotation task. In each round, we would select a batch of around 200 posts. Each post was annotated by three annotators, then majority voting was applied to decide the final label. The final dataset is composed of 9,055 posts from Twitter and 11,093 posts from Gab. Table 3 provides further details about the dataset collected. Table 4 shows samples of our dataset. The Krippendorff's α for the inter-annotator agreement is 0.46 which is much higher than other hate speech datasets (Vigna et al. 2017; Ousidhoum et al. 2019).

Ground Truth Attention

- converting each rationale into an attention vector.
  - a Boolean vector with length equal to the number of tokens in the sentence.

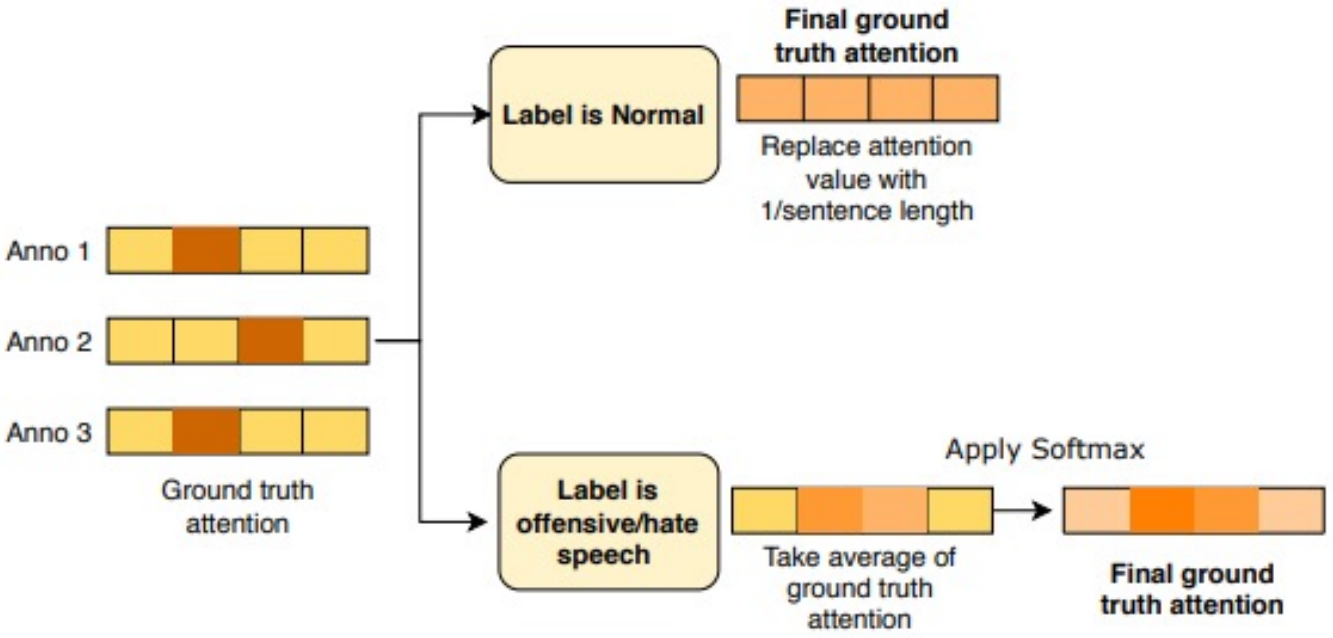


Figure 1: Ground truth attention.

Evaluation Metrics

- Performance Based Metrics:
    - accuracy, macro-F1, AUROC
- Bias Based metrics:
    - Subgroup AUC, Background Positive Subgroup Negative (BPSN) AUC, Background Negative Subgroup Positive (BNSP) AUC, Generalized Mean of Bias AUCs

Explainability Based Metrics

- Plausibility refers to how convincing the interpretation is to humans,
- Faithfulness refers to how accurately it reflects the true reasoning process of the model (Jacovi and Goldberg 2020).
    - Comprehensiveness: A high value of comprehensiveness implies that the rationales were influential in the prediction.
    - Sufficiency measures the degree to which extracted rationales are adequate for a model to make a prediction
- Agreement with human rationales.

Model [Token Method]	Performance			Explainability				
				Plausibility			Faithfulness	
	Acc.↑	Macro F1↑	AUROC↑	IOU F1↑	Token F1↑	AUPRC↑	Comp.↑	Suff.↓
CNN-GRU [LIME]	0.627	0.606	0.793	0.167	0.385	0.648	0.316	**-0.082**
BiRNN [LIME]	0.595	0.575	0.767	0.162	0.361	0.605	0.421	-0.051
BiRNN-Attn [Attn]	0.621	0.614	0.795	0.167	0.369	0.643	0.278	0.001
BiRNN-Attn [LIME]	0.621	0.614	0.795	0.162	0.386	0.650	0.308	-0.075
BiRNN-**HateXplain** [Attn]	0.629	0.629	0.805	**0.222**	**0.506**	**0.841**	0.281	0.039
BiRNN-**HateXplain** [LIME]	0.629	0.629	0.805	0.174	0.407	0.685	0.343	-0.075
BERT [Attn]	0.690	0.674	0.843	0.130	0.497	0.778	0.447	0.057
BERT [LIME]	0.690	0.674	0.843	0.118	0.468	0.747	0.436	0.008
BERT-**HateXplain** [Attn]	**0.698**	**0.687**	**0.851**	0.120	0.411	0.626	0.424	0.160
BERT-**HateXplain** [LIME]	**0.698**	**0.687**	**0.851**	0.112	0.452	0.722	**0.500**	0.004

Table 5: Model performance results. To select thon, we used attention and LIME methods.

- Models that perform very well in classification cannot always provide plausible and faithful rationales for their decisions.

Rationales predicted by different models compared to human annotators

Model	Text												Label
Human Annotator	The	[redacted]	are again using	[redacted]	as an excuse to **spread their agenda**	[redacted]	should have **eradicated** them	HS
CNN-GRU	The	[redacted]	are again using	[redacted]	as an excuse to spread **their agenda**	[redacted]	should have eradicated them	HS
BiRNN	The	[redacted]	are again using	[redacted]	as an excuse to spread their **agenda** .	[redacted]	should have eradicated them	HS
BiRNN-Attn	The	[redacted]	are again using	[redacted]	as an excuse to spread **their** agenda .	[redacted]	should have eradicated them	HS
BiRNN-HateXplain	The	[redacted]	are again using	[redacted]	as an excuse to spread **their** agenda .	[redacted]	should have **eradicated** them	HS
BERT	The	[redacted]	are again using	[redacted]	as an excuse to **spread** their **agenda**	[redacted]	should have **eradicated** them	OF
BERT-HateXplain	The	[redacted]	are again using	[redacted]	as an excuse to spread their **agenda** .	[redacted]	should have eradicated them	OF

Table 1: Example of the rationales predicted by different models compared to human annotators. The bold part marks tokens that the human annotator and model found important for the prediction. The underlined part marks tokens which the model found important, but the human annotators did not.

- even when the model is making the correct prediction, rationales varies across models.
- In case of BERT, it attends to several of the tokens that human annotators deemed important, but assigns the wrong label

[redacted] *Removed due to hate speech content, mostly marked as rationale*

Study-2:

Necessity and Sufficiency for Explaining Text Classifiers: A Case Study in Hate Speech Detection

Esma Balkır, Isar Nejadgholi, Kathleen C. Fraser, and Svetlana Kiritchenko
National Research Council Canada
Ottawa, Canada
{Esma.Balkir,Isar.Nejadgholi,Kathleen.Fraser,Svetlana.Kiritchenko}@nrc-cnrc.gc.ca

https://arxiv.org/abs/2205.03302

Methods for Explanations

What to Explain

- Explaining individual prediction (local)
- Explaining model reasoning as a whole (global)

When to generate explanations

- As a part of the prediction process (self-explaining) or
- generated through additional post-processing (post-hoc)

Methods used here :

Feature attribution methods (scores for each feature)

Counterfactual Generation Models (contrastive cases)

Necessity and Sufficiency

- Providing scores for each feature according to *how important that feature* is for the classifier decision for a given input.

- A feature that is ‘important’ might translate to two different predictions.
  - Necessity: if the value of an important feature changes, then the prediction will change.

  - Sufficiency: as long as the feature remains, the prediction will not change.

  E.g., “I hate women.”

  - “woman” token here might have high necessity and low sufficiency
    - “I hate oranges” >> ”women” token is replaced with another object, then no hate
    - “I admire woman” >> the token is still there, but it is not hate speech any more

Aim

- Applying explainability metrics to test classifiers' fairness towards identity-based groups (e.g., women, Muslims).

- Showing how necessity and sufficiency metrics calculated for identity terms over hateful sentences can explain the classifier's behavior on non-hateful statements,

Remember LIME

- It removes the items from the sentence and check whether this changes the classification!
- But removing/masking items makes the rest of the sentence unnatural!
- They propose to replace tokens with most probable n–gram prediction.

Experimental Details

- training six BERT classifiers on three different datasets (Founta et al., 2018; Vidgen et al., 2021; Davidson et al., 2017), and obtaining the necessity and sufficiency of the identity terms on explicitly hateful test cases targeting women and Muslims.

- Two training task : (i) on ‘hate speech’ labels and (ii) on more general ‘abusive’ language labels

- The necessity and sufficiency scores are only calculated for test cases that a classifier returns a correct prediction!

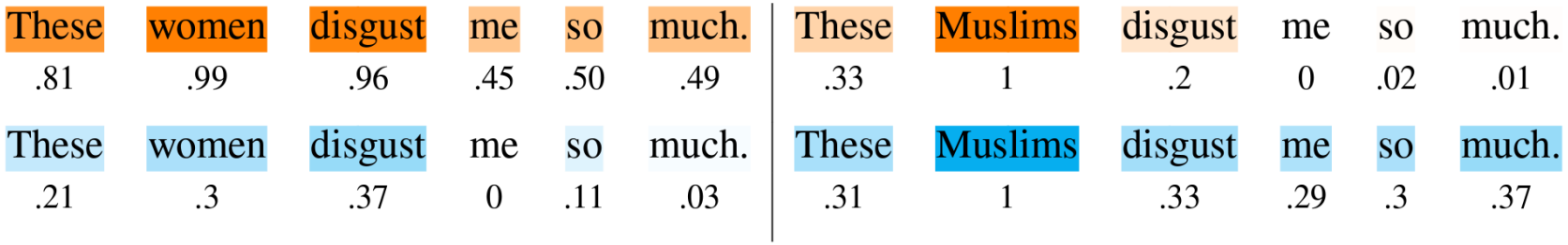


Figure 2: Visualization for necessity and sufficiency scores on an example HateCheck test case targeting *women* and *Muslims* for the classifier *Founta2018-hate*. Darker shades correspond to higher values.

- The token 'Muslims' is more sufficient compared to 'women',
- The token 'disgust 'is more necessary in the context of 'women' than that of 'Muslims'.
- High sufficiency means over-sensitivity to identity terms.

  >>> the mere occurrence of the word "Muslims" is sufficient for the classifiers to classify a text as hate speech, even if the text is neutral

  >>> the highest error rates on these test cases

Results

- As baselines, the average importance of the tokens corresponding to target groups were calculated with SHAP and LIME. LIME and SHAP scores are inconsistent in detecting bias.

- But, theoretically-grounded concepts of necessity and sufficiency are better at explaining the classifiers.

Back to Implicit Hate Speech

A Study In Hate: Dissecting Transformer-Based Models' Rationale for Implicit Hate Classification

Faye Holt and **Cuong Nguyen** and **Parth Shah**
Georgia Institute of Technology

Dataset and Model

- On the Implicit Hate Speech Dataset (a smaller subset)

- Models:
  - SVM as baseline classifier
  - BERT (masked language model)
  - ELECTRA: Instead of corrupting the input by replacing tokens with “[MASK]” as in BERT, ELECTRA corrupts the input by replacing some input tokens with incorrect, but somewhat plausible, fakes.
  - DistillBERT: a distilled version of BERT, smaller, faster, cheaper and lighter"

Hate & No-Hate Classification

Label	SVM (baseline)	BERT	ELECTRA	DistilBERT
Precision	0.49	0.656	**0.737**	0.684
Recall	0.68	0.71	**0.706**	0.677
F1	0.59	0.693	**0.721**	0.68
Accuracy	0.76	0.77	**0.812**	0.794

Implicit Hate Subcategories

Label	SVM (bl) Acc	BERT Acc
0 Stereotypical	0.62	**0.73**
1 White Grievance	**0.63**	0.14
2 Incitement	**0.5**	0.18
3 Threatening	**0.38**	0.25
4 Other	0	0
5 Inferiority	**0.14**	0
6 Irony	**0.74**	0.68
F1	**0.43**	0.28

Table 6: Multi-Class Model Results (Accuracy and F1)

SHAP Values as explanations

- SHAP values, a game-theoretic concept that intuitively describes each feature's contribution to the final outcome, after taking into account all possible combinations of features.

- SHAP feature scores can not only be calculated for localized samples but also for the entire global dataset

- Different Shapley metrics for different models:
  - kernelSHAP (used for SVM),
  - deepSHAP (used for BERT).

Effect of Text Preprocessing

Removal Strategy	F1
Base (No Removal)	**0.64**
Punctuations	0.6240
Stopwords	0.6229
Twitter Artefacts (TA)	0.6156
Punctuations + Stopwords + Twitter Artefacts	0.5911

Table 8: Results for Ablation Test

Punctuation	BERT	ELECTRA	DistilBERT
!	0.4006	2.149	1.4377
"	0.2647	1.713	0.4288
#	0.2413	0.6521	0.3909
’	2.1152	2.6717	0.8901
(	0.0883	0.4376	0.0548
)	0.0197	0.0503	0.0399
,	0.7749	0.6413	0.7259
.	1.2342	2.1705	1.6963
/	0.3822	1.32	0.1336
?	0.9752	1.8565	3.6899
@	0.531	1.8426	0.6023

Table 9: Maximal SHAP Values for selected punctuations (with logit scaling)

BERT Binary	BERT Multi-class					
Hate	Irony	Stereotypical	White Grievance	Incitement	Threatening	Inferiority
wake	getting	wants	–	–	–	wants
whites	tile	getting	wants	getting	brotherhood	tman
rape	**islam**	**ukraine**	change	violent	violent	utation
sell	gen	com	etc	**ukraine**	com	call
je	ev	gen	come	–	person	p
islamic	etc	per	crime	com	pen	thirty-one
try	booklet	crime	com	holm	forty-seven	ong
amnesty	wants	hurting	groups	wants	tile	tion
white	**ukraine**	make	murders	forty-seven	nazis	come
jewish	come	thirty-one	solidarity	support	holm	holm
black	nazis	festival	tman	shoved	words	**ukraine**
council	festival	attending	network	per	american	king
kill	**quran**	come	words	sign	**jews**	screwed
alien	**jewish**	**islam**	gen	crime	biggest	tics

Table 7: Top 15 weighted features for binary and multi-class models. Highlights indicate similar features across binary and multi-class. Bolded words indicate features chosen to mask.

- Our models are learning that sentences with identity based nouns are the main indicator of hate speech.

Take-away messages

- Explanations are insightful to have a better understanding of what is happening in the black-box models, as well as their fairness and bias
- There are different methods to extract explanations.
- There are different metrics to evaluate the explanations
- Ablation study + explanations
- Current classifiers, even deep contextualized ones, seem to rely on the lexical presence of identity terms and ignore the characteristics of the object of abuse

Project-I:
Hate Sense Classification

- Needs >= 2 students, preferably 4 (for interrater agreement and sound evaluation)

	Annotator-1	~~**Annotator-3**~~	~~**Annotator-3**~~	~~**Annotator-4**~~
Task-1	500 sentence (batch-1)	~~500 sentence (batch-1)~~	~~500 sentence (batch-2)~~	~~500 sentence (batch-2)~~
Task-2	~~Interrater agreement~~		~~Interrater agreement~~	
Task-3	**Train on 1st part, test on 2nd part (using the provided model)**		**Train on 2nd part, test on 1st part (using the provided model)**	
Task-4	Report the results (discuss annotation scheme)			

Project-II:
Implicit versus Explicit Hate Speech (Gaze4Hate)

- Needs >= 2 students (but we also have gold label annotations)
- Your tasks:

	Annotator-1	Annotator-2
~~**Task-1**~~	~~90 sentences (implicit/explicit/other)~~	~~90 sentences (implicit/explicit/other)~~
~~**Task-2**~~	~~Interrater agreement~~	
~~**Task-3**~~	~~CHANGED: select a list of models from Huggingface~~	
Task-4	**use the provided ML Classifier**	
Task-5	**Report the results,**	
Task-6	Do linguistic analysis between explicit and implicit cases	

Project-IV: Explainability

- Reference Dataset: HateCheck or HateCheck Multilingual
- (https://hatecheck.ai/download/)
- ML Task: given a sentence, decide whether the sentence is hateful or not and extract explanations,
- Your task:
  - ~~Choose a test dataset (that exists in Hate Check or HateCheck Multilingual)~~
  - ~~Implement an existing pretrained hate speech model from the Huggingface platform~~
  - ~~Optional: finetune it on another suitable dataset~~
  - **extract explanations using existing python libraries e.g. Captum, LIME etc.**
  - Interpret the results with the functional categories provided in HateCheck templates
  - Report the results

Project-IV:
Curriculum Learning Training

Task-1	~~Choose your dataset e.g. from multilingual HateCheck~~
Task-2	~~Manually inspect the existing groups and define some complexity criteria~~
Task-3	~~Create n complexity groups, and rank the instances w.r.t. the criteria~~
Task-4	**Implement an existing pretrained hatespeech model from the Huggingface platform and fine tune it step-by-step on the reranked training data**
Task-5	**Optional: conduct reranking procedure based on training loss on each instance**
Task-6	Report the results
Task-7	Discuss possible curriculums