Hate Speech Detection *WiSe 23-24*
Contextualized Embeddings

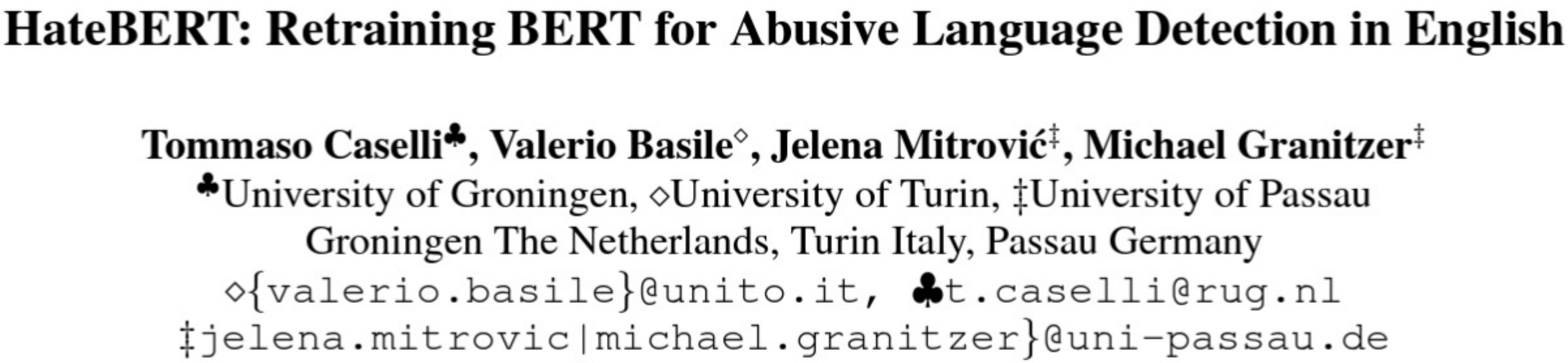
Image taken from https://deepsense.ai/artificial-intelligence-hate-speech/

(Dr.) Özge Alaçam

Computational Linguistics

oezge.alacam@lmu.de

Why don’t we prefer to use the BERT base model on hate speech classification?



HateBERT: Retraining BERT for Abusive Language Detection in English

Tommaso Caselli[♣], Valerio Basile[◇], Jelena Mitrović[‡], Michael Granitzer[‡]
♣University of Groningen, ◇University of Turin, ‡University of Passau
Groningen The Netherlands, Turin Italy, Passau Germany
◇{valerio.basile}@unito.it, ♣t.caselli@rug.nl
‡jelena.mitrovic|michael.granitzer}@uni-passau.de

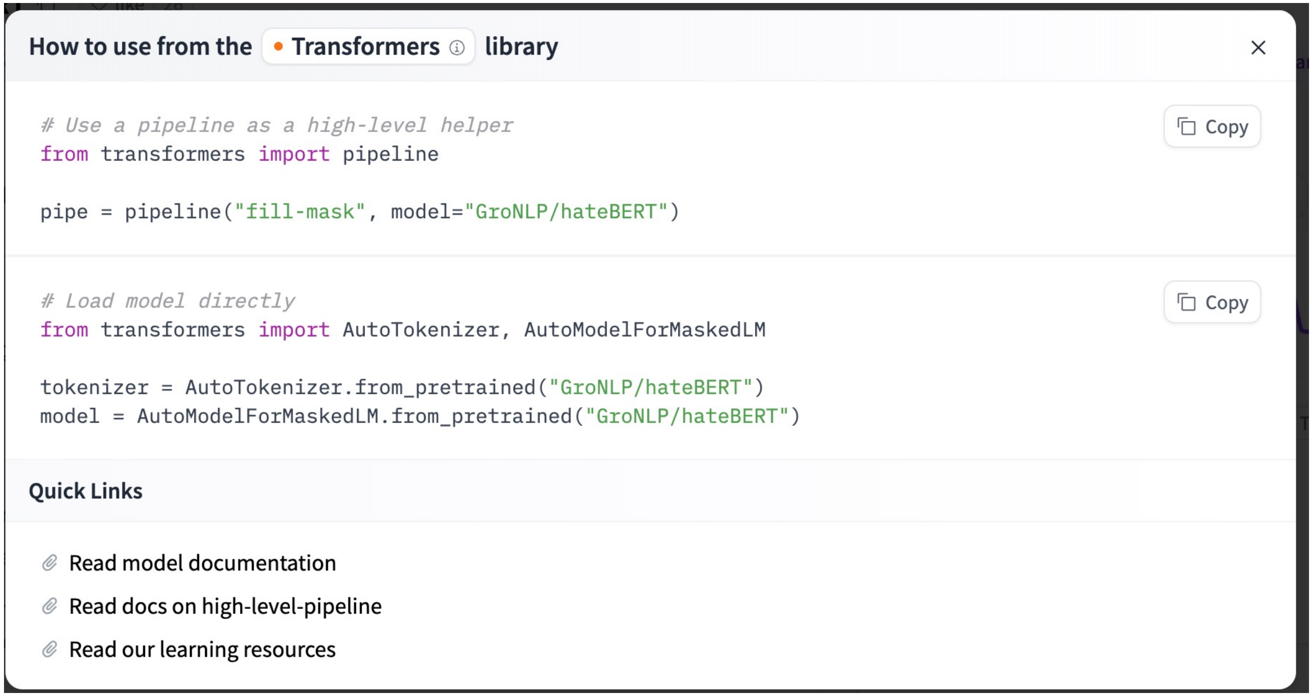
- Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2020). Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

- Huggingface Platform:

- https://huggingface.co/GroNLP/hateBERT

- HateBERT is an English pre-trained BERT model obtained by further training the English BERT base uncased model with more than 1 million posts (RAL-E dataset) from banned communites from Reddit.

You can access this code from the Huggingface platform (under "</> Use in Transformers"



How to use from the • Transformers ⓘ library ×

```
# Use a pipeline as a high-level helper
from transformers import pipeline

pipe = pipeline("fill-mask", model="GroNLP/hateBERT")
```

```
# Load model directly
from transformers import AutoTokenizer, AutoModelForMaskedLM

tokenizer = AutoTokenizer.from_pretrained("GroNLP/hateBERT")
model = AutoModelForMaskedLM.from_pretrained("GroNLP/hateBERT")
```

Quick Links

- Read model documentation
- Read docs on high-level-pipeline
- Read our learning resources

Key contributions

- additional evidence that further pre-training is a viable strategy to obtain domain-specific in a fast and cheap way
- the release of HateBERT, a pre-trained BERT for abusive language phenomena
- the release of a large-scale dataset of social media posts in English from communities banned for being offensive, abusive, or hateful (RAL-E collected from REDDIT)
    - 1,492,740 messages from a period between 2012 and 2015, for a total of 43,820,621 tokens

Creating HateBERT

- From the RAL-E dataset, they used 1,478,348 messages to re-train the English BERT base-uncased model by applying the Masked Language Model (MLM) objective.
  - Re-trained for 100 epochs (almost 2 million steps) in batches of 64 samples, including up to 512 sentence piece tokens.
  - Adam with learning rate 5e-5.
  - using the huggingface code on one Nvidia V100 GPU.
- The remaining 15K messages have been used as test set.
- The result is a shifted BERT model, HateBERT base-uncased, along two dimensions: (i.) language variety (i.e. social media); and (ii.) polarity (i.e., offense-, abuse-, and hate-oriented model).

Pre-processing before re-training

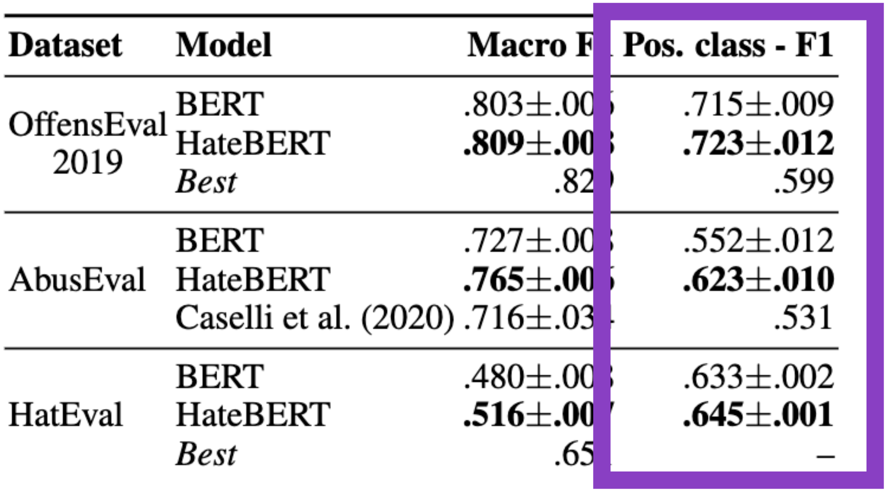
- all users’ mentions have been substituted with a placeholder (@USER);
- all URLs have been substituted with a with a placeholder (URL);
- emojis have been replaced with text (e.g. → :pleading face:) using Python emoji package;
- hashtag symbol has been removed from hasthtags (e.g. #kadiricinadalet → kadiricinadalet);
- extra blank spaces have been replaced with a single space;
- extra blank new lines have been removed.

Pre-processing before fine-tuning

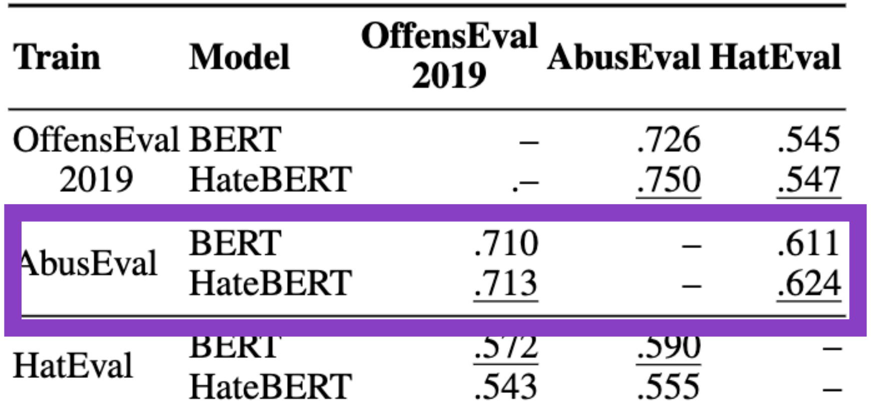
- all users’ mentions have been substituted with a placeholder (@USER);
- all URLs have been substituted with a with a placeholder (URL);
- emojis have been replaced with text (e.g. → :pleading face:) using Python emoji package;
- hashtag symbol has been removed from hashtags (e.g. #kadiricinadalet → kadiricinadalet);
- extra blank spaces have been replaced with a single space.

Testing on

- OffensEval 2019 (Zampieri et al., 2019)
- AbusEval (Caselli et al., 2020)
- HatEval (Basile et al., 2019)



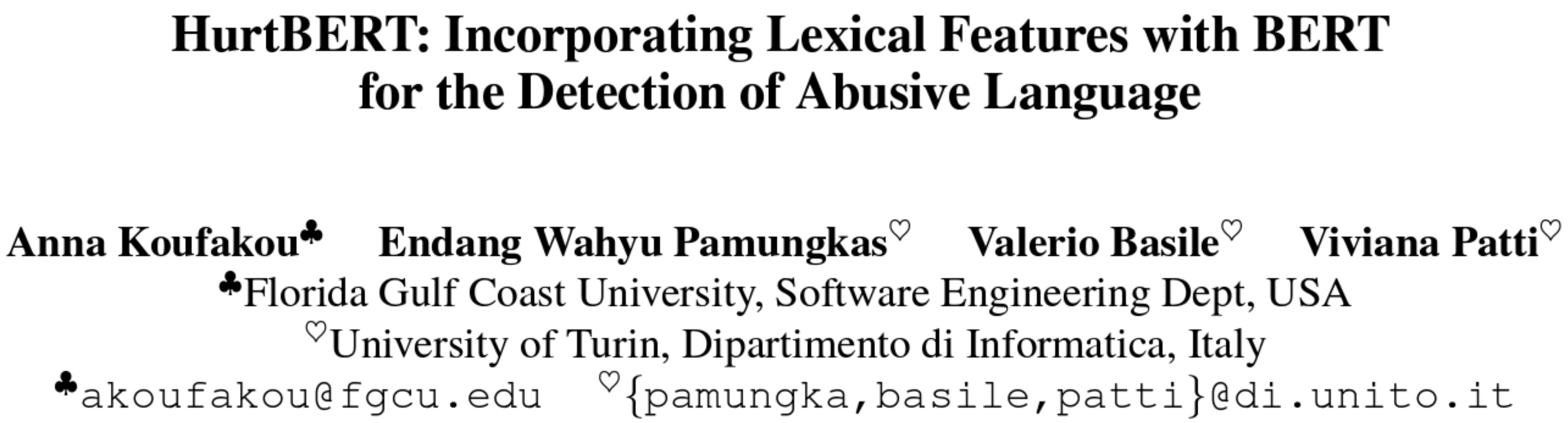
Dataset	Model	Macro F	Pos. class - F1
OffensEval 2019	BERT	.803±.00	.715±.009
	HateBERT	**.809±.00**	**.723±.012**
	Best	.82	.599
AbusEval	BERT	.727±.00	.552±.012
	HateBERT	**.765±.00**	**.623±.010**
	Caselli et al. (2020)	.716±.03	.531
HatEval	BERT	.480±.00	.633±.002
	HateBERT	**.516±.00**	**.645±.001**
	Best	.65	–



Train	Model	OffensEval 2019	AbusEval	HatEval
OffensEval 2019	BERT	–	.726	.545
	HateBERT	.–	.750	.547
AbusEval	BERT	.710	–	.611
	HateBERT	.713	–	.624
HatEval	BERT	.572	.590	–
	HateBERT	.543	.555	–

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ENSEMBLE MODELS with/out BERT embeddings



HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language

Anna Koufakou[♣] Endang Wahyu Pamungkas[♡] Valerio Basile[♡] Viviana Patti[♡]
[♣]Florida Gulf Coast University, Software Engineering Dept, USA
[♡]University of Turin, Dipartimento di Informatica, Italy
[♣]akoufakou@fgcu.edu [♡]{pamungka,basile,patti}@di.unito.it

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 34–43
Online, November 20, 2020. c© 2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

State-of-the-art method

- BERT (Bidirectional Encoder Representations from **Transformers**)
  - state of the art in several NLP tasks, including abusive and offensive language detection
  - in the SemEval 2019 Task 6 (Zampieri et al., 2019b, OffensEval), seven out of the top-ten teams used BERT.
  - *the trend is the same for the later events as well!*

De-facto approach

- pre-training on a large quantity of text,
- then fine-tuning to a specific dataset in order to learn complex correlations between the natural language and the labels
- It does not require intensive feature engineering and learns implicit knowledge (including syntactic, semantic and discourse-level information)
- no additional external knowledge is taken into consideration, such as linguistic information from a lexicon.

HurtBERT – a hybrid approach

- infusing external knowledge into a supervised model for abusive language detection.

- HurtLex (Bassignana et al., 2018), a multilingual lexicon of offensive words, created by semi-automatically translating a handcrafted resource in Italian into 53 languages.

- A pre-trained BERT model

- two ways of extracting HurtLex features: encodings and embeddings.

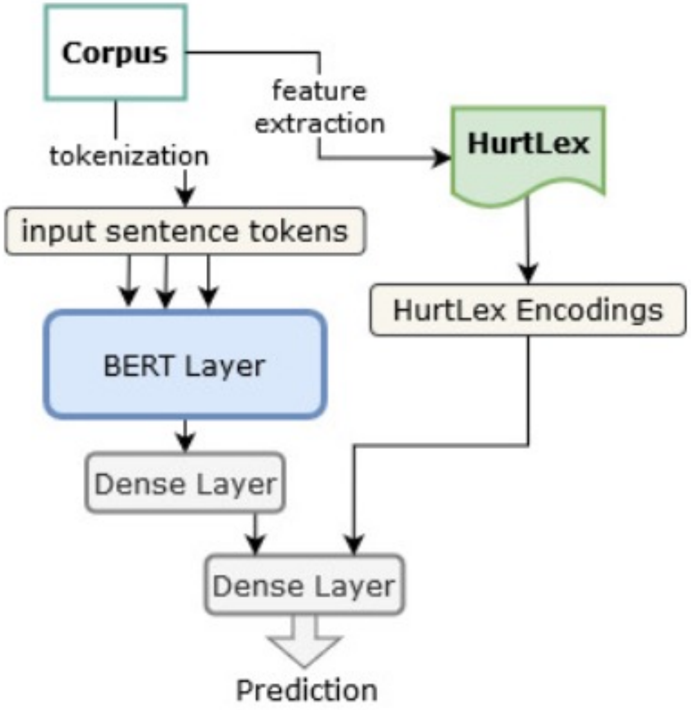


Figure 1: HurtBERT-Enc, our model using HurtLex Encodings

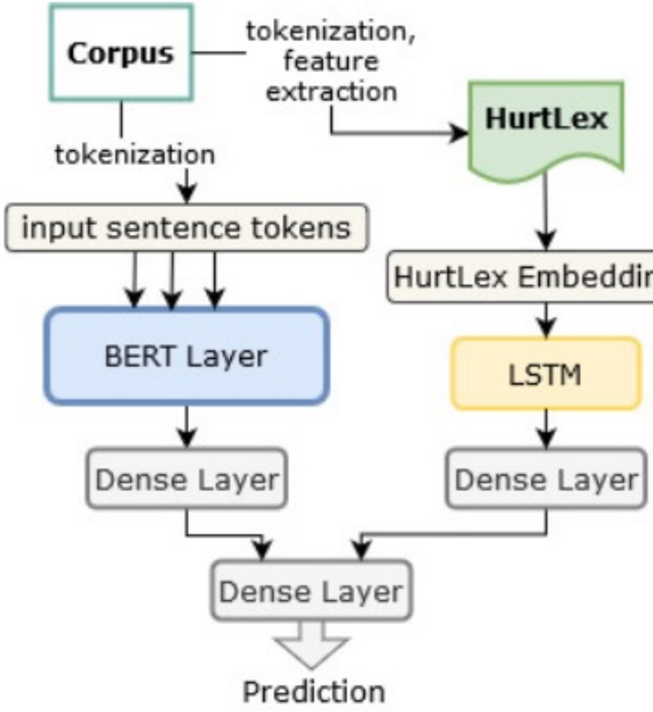


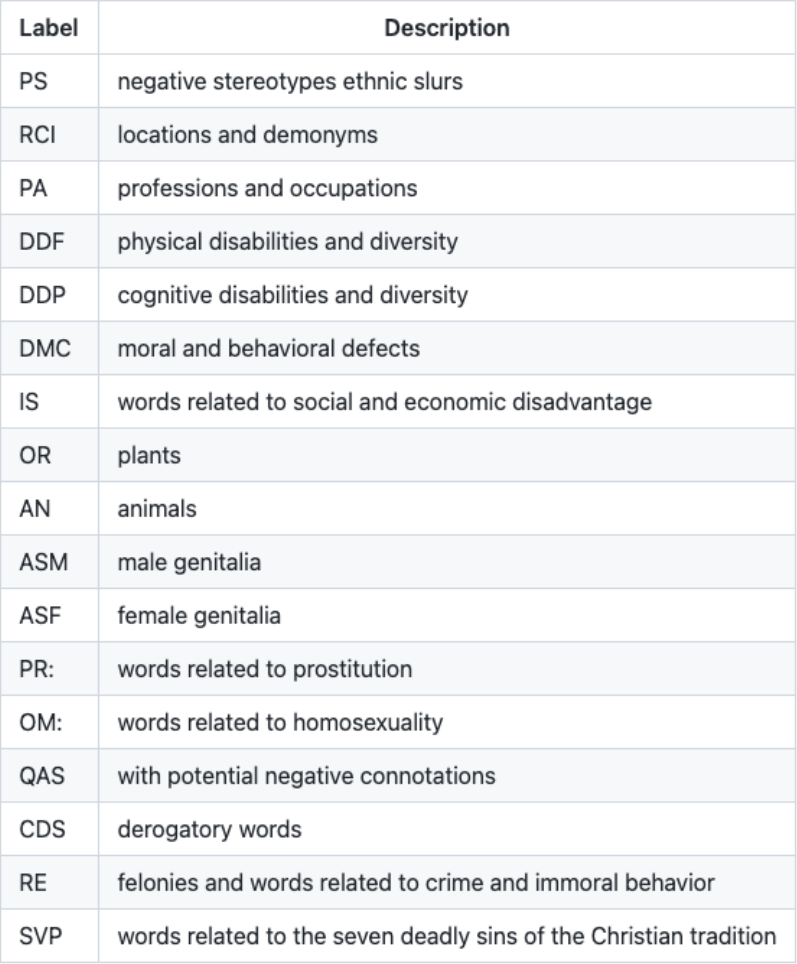
Figure 2: HurtBERT-Emb, our model using HurtLex Embeddings

Model-1

- Standart BERT uncased model

Method-2: HurtLex Encoding

- For each word in each text, check their categories in HurtLex and create a vector
- 17 categories in HurtLex >> dimensionality of the HurtLex encoding is 17.
- Each element in this vector is simply a frequency count.
- Each comment has 17D vector



Label	Description
PS	negative stereotypes ethnic slurs
RCI	locations and demonyms
PA	professions and occupations
DDF	physical disabilities and diversity
DDP	cognitive disabilities and diversity
DMC	moral and behavioral defects
IS	words related to social and economic disadvantage
OR	plants
AN	animals
ASM	male genitalia
ASF	female genitalia
PR:	words related to prostitution
OM:	words related to homosexuality
QAS	with potential negative connotations
CDS	derogatory words
RE	felonies and words related to crime and immoral behavior
SVP	words related to the seven deadly sins of the Christian tradition

Method-3: HurtLex Embeddings

- The HurtLex embedding is a 17-dimension one-hot encoding of the word presence in each lexicon category.
- Created by an LSTM model (a sequence model)
- the encoding is a simple representation that reflects how many times the category is found in the text. While the embedding-based model also represents non-linear interactions between the features, that is, linguistically, the role of the HurtLex words in the sentence.
- Each word has 17D vector

A collection of datasets

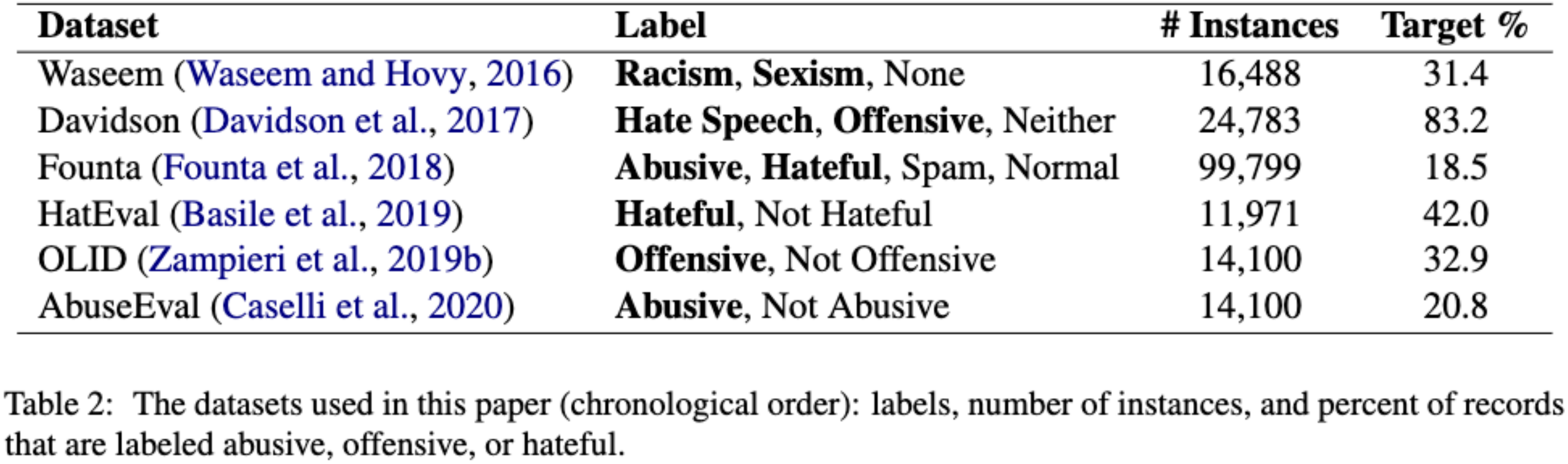
- Selection criteria: Binary labels, in English
- Split into training, development and test sets (70%, 10% and 20%)
- Waseem (Waseem and Hovy, 2016) :
  - 17K tweets, sexist (3,3K), racist (2K), and neither (11,5K)
- Davidson (Davidson et al., 2017) : 24,7K tweets,
  - hate (5.8%), offensive (77.4%), not offensive (16.8%).
- Founta (Founta et al., 2018). : 80K tweets
  - Abusive (11%), hateful (7.5%), spam (22.5%), and normal (59%)

Tip! Don't use different styles in your research paper e.g. reporting size in number versus percentage

- HatEval (Basile et al., 2019). 12K tweet
  - Against Immigrants and Women in Twitter

- OLID (Zampieri et al., 2019a) Offensive (30%) and Not Offensive labeled data, where about 30% of the records are labeled as Offensive.

- AbuseEval: Caselli et al. (2020) on implicit and explicit abusive language.



Dataset	Label	# Instances	Target %
Waseem (Waseem and Hovy, 2016)	**Racism**, **Sexism**, None	16,488	31.4
Davidson (Davidson et al., 2017)	**Hate Speech**, **Offensive**, Neither	24,783	83.2
Founta (Founta et al., 2018)	**Abusive**, **Hateful**, Spam, Normal	99,799	18.5
HatEval (Basile et al., 2019)	**Hateful**, Not Hateful	11,971	42.0
OLID (Zampieri et al., 2019b)	**Offensive**, Not Offensive	14,100	32.9
AbuseEval (Caselli et al., 2020)	**Abusive**, Not Abusive	14,100	20.8

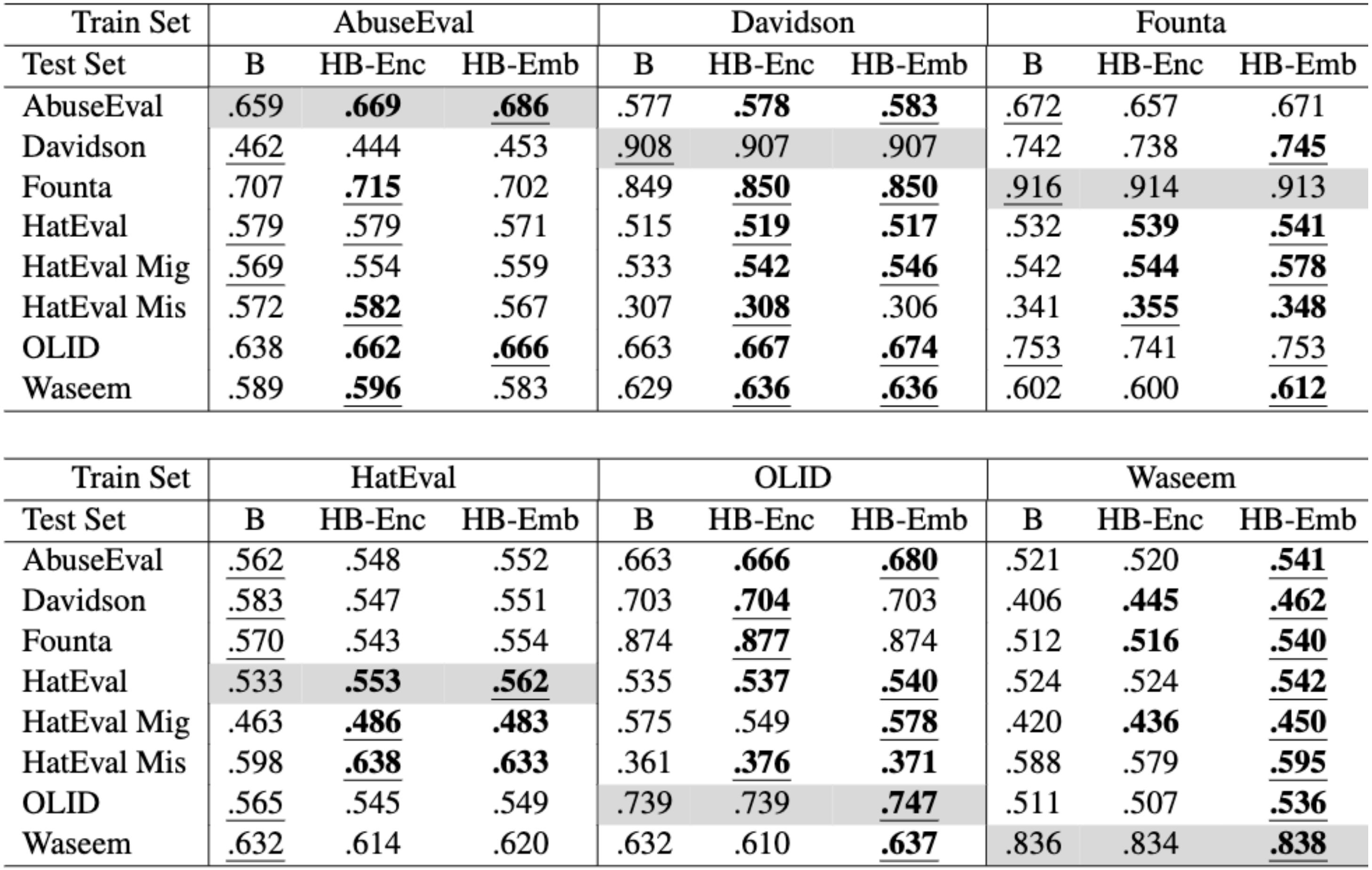
Table 2: The datasets used in this paper (chronological order): labels, number of instances, and percent of records that are labeled abusive, offensive, or hateful.

- Finally, for both models, they concatenate the dense layer from the BERT output and the dense layer from the HurtLex output, before passing into a dense layer with sigmoid activation as the predictor layer

- 6 datasets
- Training on the training set of each set and
- Then test them on each test set resulting

  720 experiments (3 models × 6 train sets × 8 test sets × 5 runs).

Results



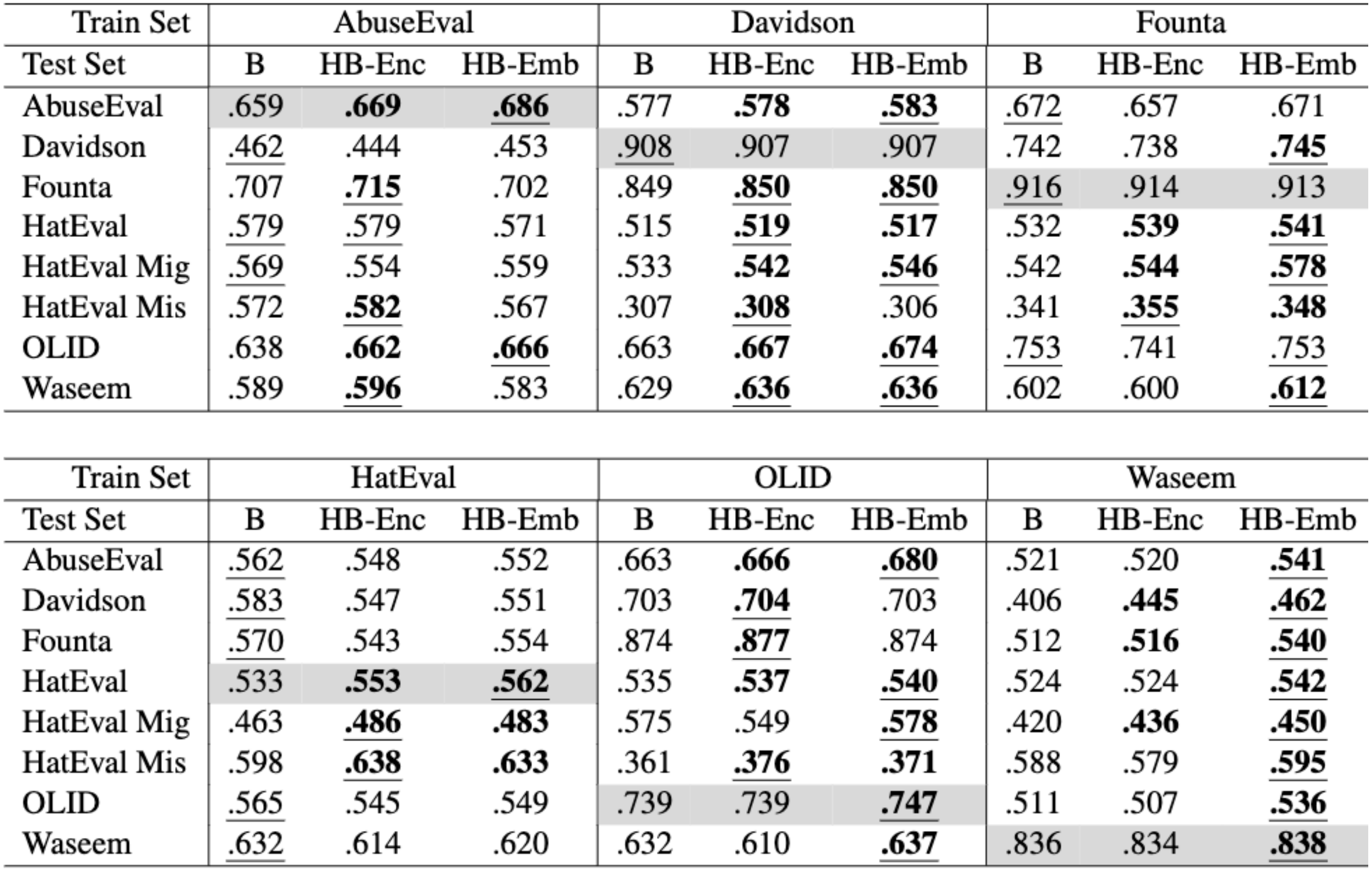
Train Set	AbuseEval			Davidson			Founta		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.659	**.669**	**.686**	.577	**.578**	**.583**	.672	.657	.671
Davidson	.462	.444	.453	.908	.907	.907	.742	.738	**.745**
Founta	.707	**.715**	.702	.849	**.850**	**.850**	.916	.914	.913
HatEval	.579	.579	.571	.515	**.519**	**.517**	.532	**.539**	**.541**
HatEval Mig	.569	.554	.559	.533	**.542**	**.546**	.542	**.544**	**.578**
HatEval Mis	.572	**.582**	.567	.307	**.308**	.306	.341	**.355**	**.348**
OLID	.638	**.662**	**.666**	.663	**.667**	**.674**	.753	.741	.753
Waseem	.589	**.596**	.583	.629	**.636**	**.636**	.602	.600	**.612**

Train Set	HatEval			OLID			Waseem		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.562	.548	.552	.663	**.666**	**.680**	.521	.520	**.541**
Davidson	.583	.547	.551	.703	**.704**	.703	.406	**.445**	**.462**
Founta	.570	.543	.554	.874	**.877**	.874	.512	**.516**	**.540**
HatEval	.533	**.553**	**.562**	.535	**.537**	**.540**	.524	.524	**.542**
HatEval Mig	.463	**.486**	**.483**	.575	.549	**.578**	.420	**.436**	**.450**
HatEval Mis	.598	**.638**	**.633**	.361	**.376**	**.371**	.588	.579	**.595**
OLID	.565	.545	.549	.739	.739	**.747**	.511	.507	**.536**
Waseem	.632	.614	.620	.632	.610	**.637**	.836	.834	**.838**

Table 3: The F1-macro results for all datasets. Shaded means in-dataset experiment. *B* stands for the baseline, *HB-Enc* stands for HurtBERT-Enc, and *HB-Emb* stands for HurtBERT-Emb. Bold indicates our model improves on the baseline; underlined indicates the best result (max). Each result is the average of five runs.

HurtBERT performs better than the baseline on 4 out of 6 datasets, namely AbuseEval, HatEval, OLID, and Waseem

Results



Train Set	AbuseEval			Davidson			Founta		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.659	**.669**	**.686**	.577	**.578**	**.583**	.672	.657	.671
Davidson	.462	.444	.453	.908	.907	.907	.742	.738	**.745**
Founta	.707	**.715**	.702	.849	**.850**	**.850**	.916	.914	.913
HatEval	.579	.579	.571	.515	**.519**	**.517**	.532	**.539**	**.541**
HatEval Mig	.569	.554	.559	.533	**.542**	**.546**	.542	**.544**	**.578**
HatEval Mis	.572	**.582**	.567	.307	**.308**	.306	.341	**.355**	**.348**
OLID	.638	**.662**	**.666**	.663	**.667**	**.674**	.753	.741	.753
Waseem	.589	**.596**	.583	.629	**.636**	**.636**	.602	.600	**.612**

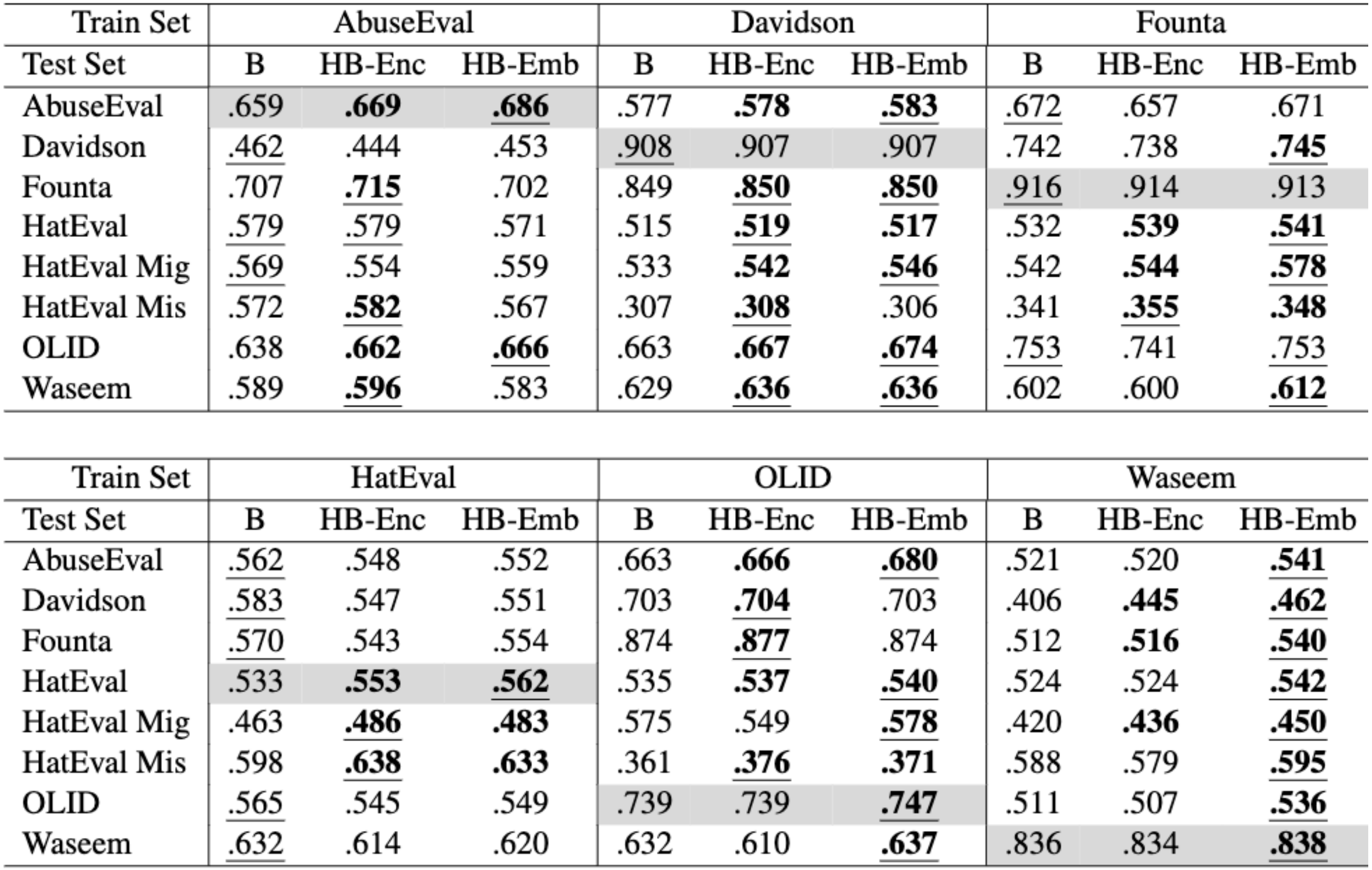
Train Set	HatEval			OLID			Waseem		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.562	.548	.552	.663	**.666**	**.680**	.521	.520	**.541**
Davidson	.583	.547	.551	.703	**.704**	.703	.406	**.445**	**.462**
Founta	.570	.543	.554	.874	**.877**	.874	.512	**.516**	**.540**
HatEval	.533	**.553**	**.562**	.535	**.537**	**.540**	.524	.524	**.542**
HatEval Mig	.463	**.486**	**.483**	.575	.549	**.578**	.420	**.436**	**.450**
HatEval Mis	.598	**.638**	**.633**	.361	**.376**	**.371**	.588	.579	**.595**
OLID	.565	.545	.549	.739	.739	**.747**	.511	.507	**.536**
Waseem	.632	.614	.620	.632	.610	**.637**	.836	.834	**.838**

Table 3: The F1-macro results for all datasets. Shaded means in-dataset experiment. *B* stands for the baseline, *HB-Enc* stands for HurtBERT-Enc, and *HB-Emb* stands for HurtBERT-Emb. Bold indicates our model improves on the baseline; underlined indicates the best result (max). Each result is the average of five runs.

HurtBERT performs better than the baseline on 4 out of 6 datasets, namely AbuseEval, HatEval, OLID, and Waseem.

In all four cases, HurtBERT-Emb is doing the best.

Results



Train Set	AbuseEval			Davidson			Founta		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.659	**.669**	**.686**	.577	**.578**	**.583**	.672	.657	.671
Davidson	.462	.444	.453	.908	.907	.907	.742	.738	**.745**
Founta	.707	**.715**	.702	.849	**.850**	**.850**	.916	.914	.913
HatEval	.579	.579	.571	.515	**.519**	**.517**	.532	**.539**	**.541**
HatEval Mig	.569	.554	.559	.533	**.542**	**.546**	.542	**.544**	**.578**
HatEval Mis	.572	**.582**	.567	.307	**.308**	.306	.341	**.355**	**.348**
OLID	.638	**.662**	**.666**	.663	**.667**	**.674**	.753	.741	.753
Waseem	.589	**.596**	.583	.629	**.636**	**.636**	.602	.600	**.612**

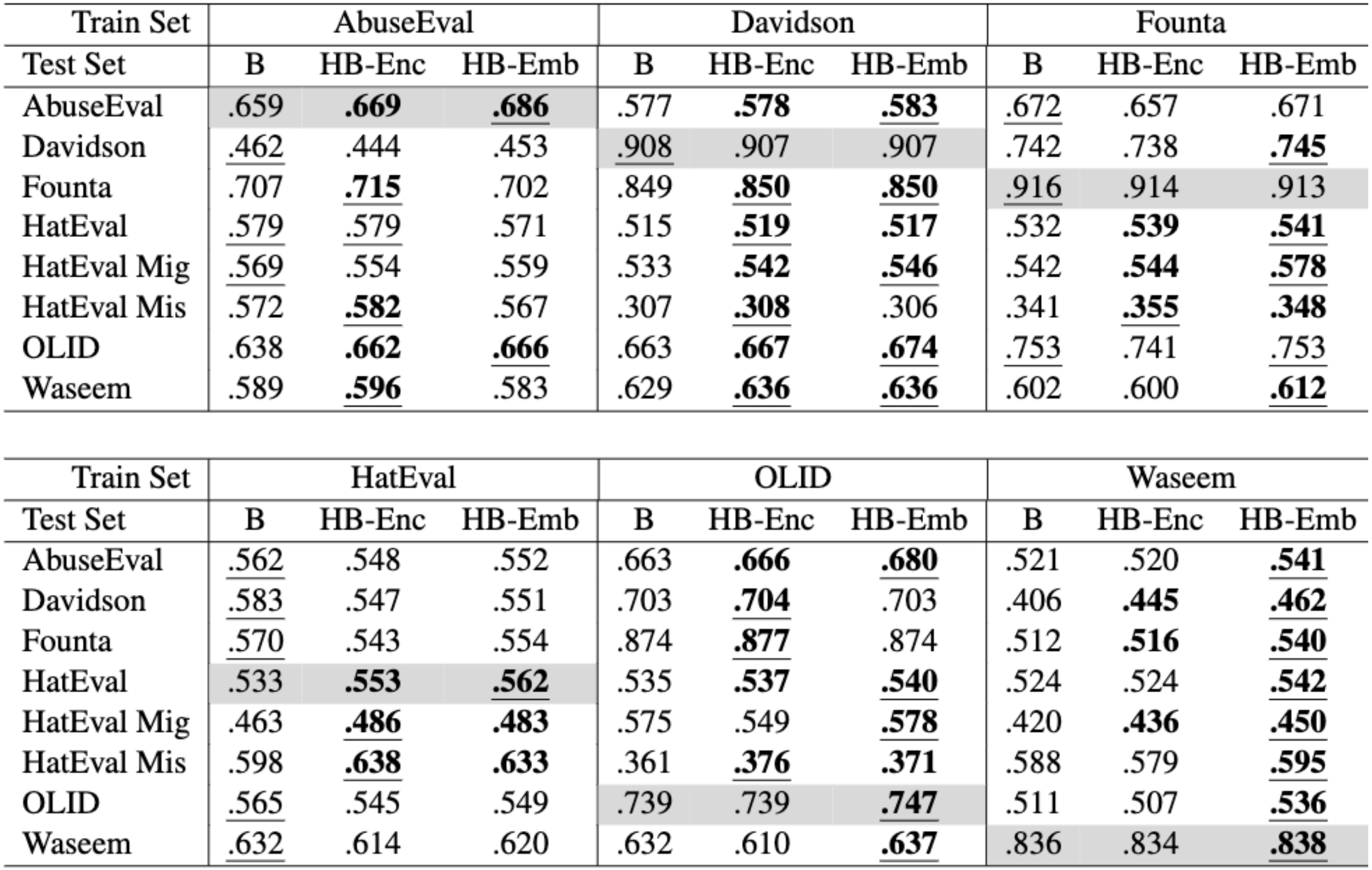
Train Set	HatEval			OLID			Waseem		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.562	.548	.552	.663	**.666**	**.680**	.521	.520	**.541**
Davidson	.583	.547	.551	.703	**.704**	.703	.406	**.445**	**.462**
Founta	.570	.543	.554	.874	**.877**	.874	.512	**.516**	**.540**
HatEval	.533	**.553**	**.562**	.535	**.537**	**.540**	.524	.524	**.542**
HatEval Mig	.463	**.486**	**.483**	.575	.549	**.578**	.420	**.436**	**.450**
HatEval Mis	.598	**.638**	**.633**	.361	**.376**	**.371**	.588	.579	**.595**
OLID	.565	.545	.549	.739	.739	**.747**	.511	.507	**.536**
Waseem	.632	.614	.620	.632	.610	**.637**	.836	.834	**.838**

Table 3: The F1-macro results for all datasets. Shaded means in-dataset experiment. *B* stands for the baseline, *HB-Enc* stands for HurtBERT-Enc, and *HB-Emb* stands for HurtBERT-Emb. Bold indicates our model improves on the baseline; underlined indicates the best result (max). Each result is the average of five runs.

the vast majority of our out-domain results are lower than the in-domain ones.

Exception: Founta and OLID

Results



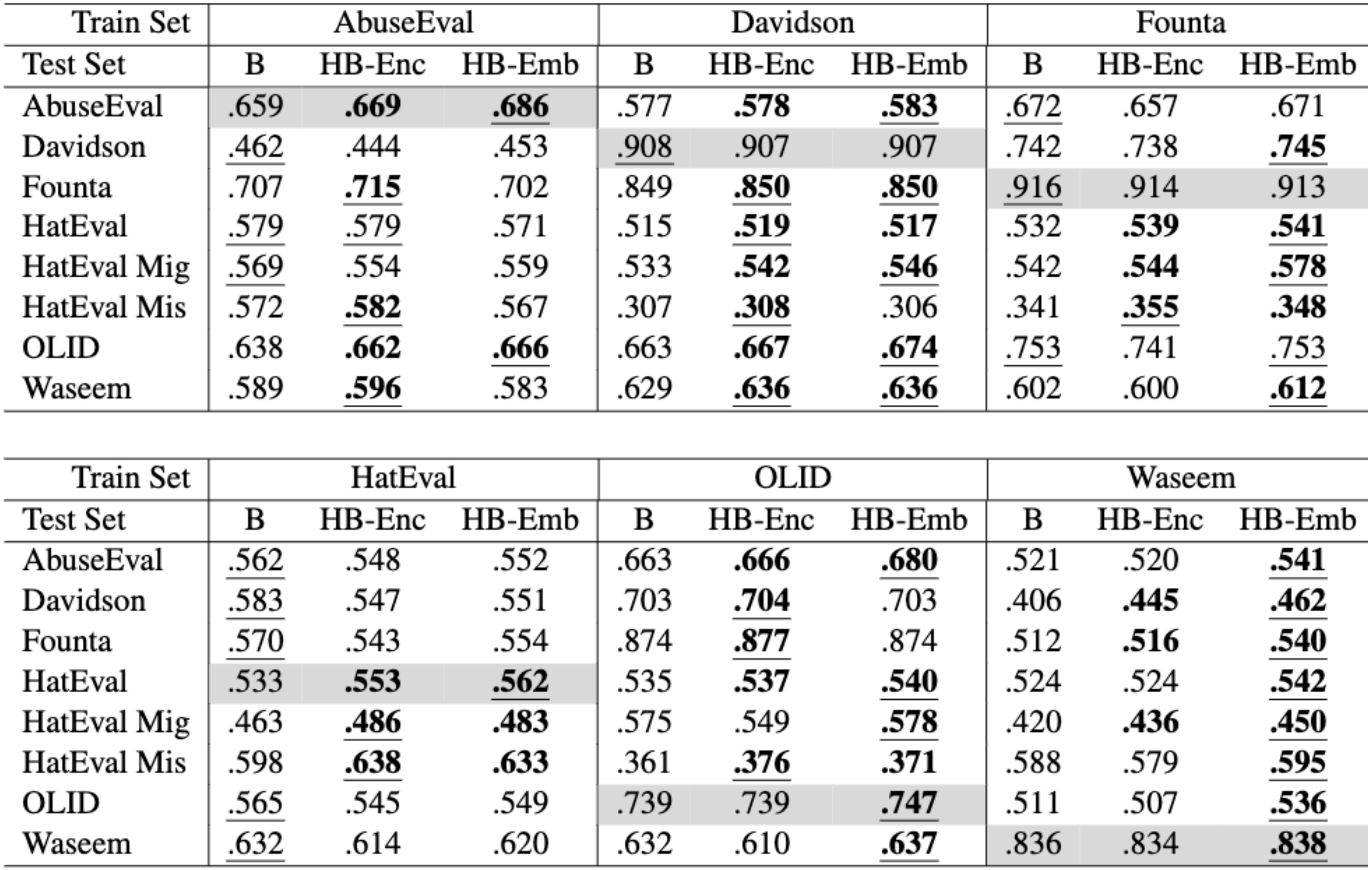
Train Set	AbuseEval			Davidson			Founta		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.659	**.669**	**.686**	.577	**.578**	**.583**	.672	.657	.671
Davidson	.462	.444	.453	.908	.907	.907	.742	.738	**.745**
Founta	.707	**.715**	.702	.849	**.850**	**.850**	.916	.914	.913
HatEval	.579	.579	.571	.515	**.519**	**.517**	.532	**.539**	**.541**
HatEval Mig	.569	.554	.559	.533	**.542**	**.546**	.542	**.544**	**.578**
HatEval Mis	.572	**.582**	.567	.307	**.308**	.306	.341	**.355**	**.348**
OLID	.638	**.662**	**.666**	.663	**.667**	**.674**	.753	.741	.753
Waseem	.589	**.596**	.583	.629	**.636**	**.636**	.602	.600	**.612**

Train Set	HatEval			OLID			Waseem		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.562	.548	.552	.663	**.666**	**.680**	.521	.520	**.541**
Davidson	.583	.547	.551	.703	**.704**	.703	.406	**.445**	**.462**
Founta	.570	.543	.554	.874	**.877**	.874	.512	**.516**	**.540**
HatEval	.533	**.553**	**.562**	.535	**.537**	**.540**	.524	.524	**.542**
HatEval Mig	.463	**.486**	**.483**	.575	.549	**.578**	.420	**.436**	**.450**
HatEval Mis	.598	**.638**	**.633**	.361	**.376**	**.371**	.588	.579	**.595**
OLID	.565	.545	.549	.739	.739	**.747**	.511	.507	**.536**
Waseem	.632	.614	.620	.632	.610	**.637**	.836	.834	**.838**

Table 3: The F1-macro results for all datasets. Shaded means in-dataset experiment. *B* stands for the baseline, *HB-Enc* stands for HurtBERT-Enc, and *HB-Emb* stands for HurtBERT-Emb. Bold indicates our model improves on the baseline; underlined indicates the best result (max). Each result is the average of five runs.

Two variants of HurtBERT obtain better results when fine-tuned on other datasets,

in particular, Davidson, OLID, and Waseem



Train Set	AbuseEval			Davidson			Founta		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.659	**.669**	**.686**	.577	**.578**	**.583**	.672	.657	.671
Davidson	.462	.444	.453	.908	.907	.907	.742	.738	**.745**
Founta	.707	**.715**	.702	.849	**.850**	**.850**	.916	.914	.913
HatEval	.579	.579	.571	.515	**.519**	**.517**	.532	**.539**	**.541**
HatEval Mig	.569	.554	.559	.533	**.542**	**.546**	.542	**.544**	**.578**
HatEval Mis	.572	**.582**	.567	.307	**.308**	.306	.341	**.355**	**.348**
OLID	.638	**.662**	**.666**	.663	**.667**	**.674**	.753	.741	.753
Waseem	.589	**.596**	.583	.629	**.636**	**.636**	.602	.600	**.612**

Train Set	HatEval			OLID			Waseem		
Test Set	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb	B	HB-Enc	HB-Emb
AbuseEval	.562	.548	.552	.663	**.666**	**.680**	.521	.520	**.541**
Davidson	.583	.547	.551	.703	**.704**	.703	.406	**.445**	**.462**
Founta	.570	.543	.554	.874	**.877**	.874	.512	**.516**	**.540**
HatEval	.533	**.553**	**.562**	.535	**.537**	**.540**	.524	.524	**.542**
HatEval Mig	.463	**.486**	**.483**	.575	.549	**.578**	.420	**.436**	**.450**
HatEval Mis	.598	**.638**	**.633**	.361	**.376**	**.371**	.588	.579	**.595**
OLID	.565	.545	.549	.739	.739	**.747**	.511	.507	**.536**
Waseem	.632	.614	.620	.632	.610	**.637**	.836	.834	**.838**

Table 3: The F1-macro results for all datasets. Shaded means in-dataset experiment. *B* stands for the baseline, *HB-Enc* stands for HurtBERT-Enc, and *HB-Emb* stands for HurtBERT-Emb. Bold indicates our model improves on the baseline; underlined indicates the best result (max). Each result is the average of five runs.

- HurtBERT-Emb has the best performance

  in 26 out of 48

  versus

- HurtBERT-Enc with *14 out of 48*

- HurtLex seems to provide more informative knowledge to the model when the goal task is to detect offensive language (e.g., OLID) rather than abusive language (e.g., AbuseEval).

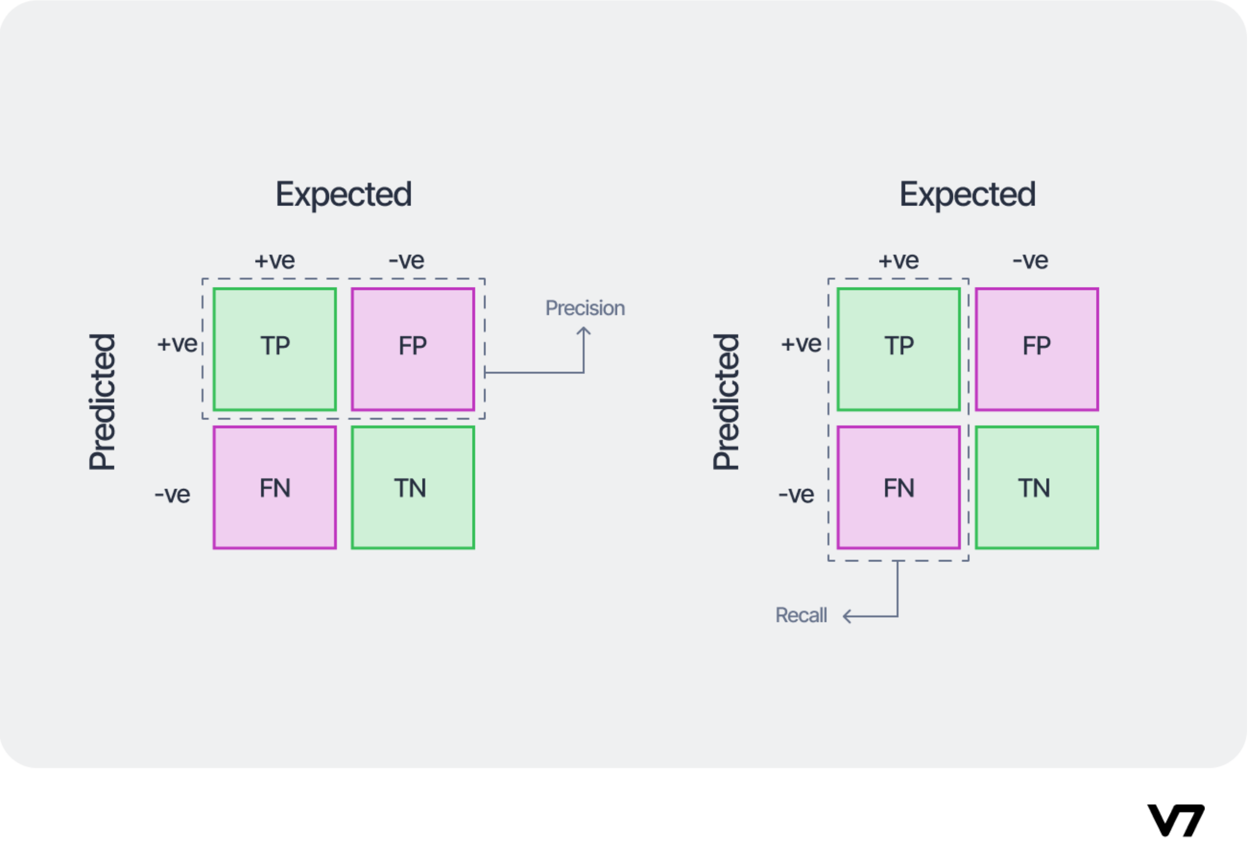
- Why?

Error Analysis

- There were many cases where swear words were present that are often used with non- offensive function.

- the additional knowledge from HurtLex has a stabilizing effect on the representation of offensive terms, whereas the fully contextual embeddings of BERT tend to always understand such terms as offensive due to the sentence- level context.

Recap (precision/recall)



Taken from https://www.v7labs.com/blog/precision-vs-recall-guide