hmBERT: Historical Multilingual Language Models for Named Entity Recognition

Stefan Schweter, Luisa März, Katharina Schmid and Erion Çano

Team HISTeria

CLEF 2022, HIPE Shared Task, 08.09.2022

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

universität wien

Software Innovation Center
DATA LAB
VOLKSWAGEN GROUP

- **Task**
- HISTeria
- Results
- Contributions

HIPE 2022 Task

Multilingual Classical Commentary Challenge

- Coarse Grained Named Entity Recognition
- NER tags: PERS, WORK, LOC, OBJECT, DATE, SCOPE
- Languages: German, French, English

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

universität wien

Software Innovation Center
DATA LAB
VOLKSWAGEN GROUP

- Task
- **HISTeria**
- Results
- Contributions

HIPE 2022 HISTeria

- hmBERT
  - Supports: German, English, French, Finnish, Swedish
  - Vocab sizes: 32k and 64k

HIPE 2022 HISTeria

hmBERT

Europeana Newspapers

- de, fr, sv, fi
- filtered for OCR confidence

Bristish Library

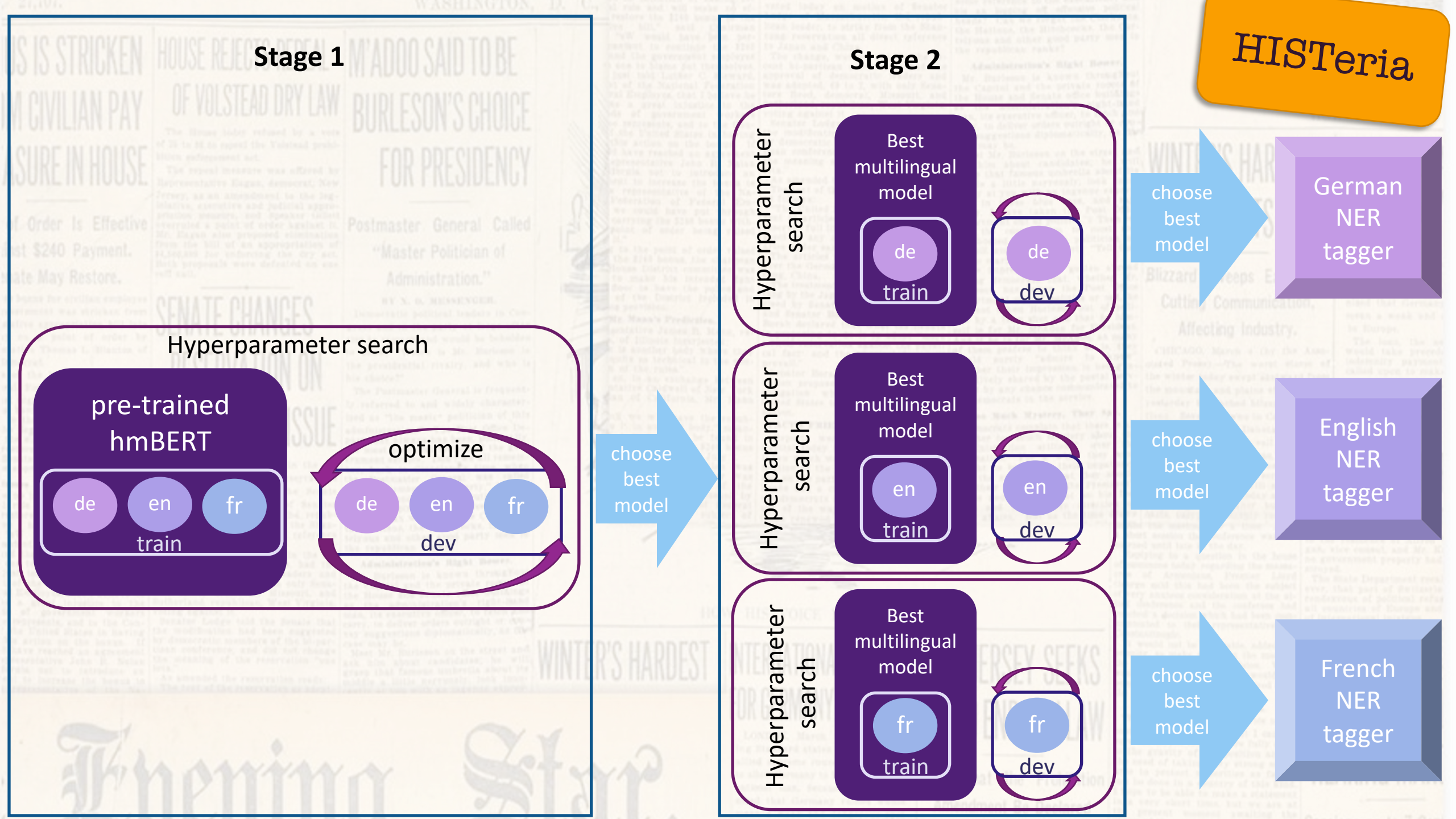
- en
- english texts between 1800 – 1900 only

10 GB per lang

Hugging Face Tokenizer cased vocab

Final corpus ~ 27 GB per language

pre-trained hmBERT


HISTeria
Stage 1
Hyperparameter search
pre-trained hmBERT
de
en
fr
train
optimize
de
en
fr
dev
choose best model
Stage 2
Hyperparameter search
Best multilingual model
de
train
de
dev
choose best model
German NER tagger
Hyperparameter search
Best multilingual model
en
train
en
dev
choose best model
English NER tagger
Hyperparameter search
Best multilingual model
fr
train
fr
dev
choose best model
French NER tagger

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

universität wien

Software Innovation Center
DATA LAB
VOLKSWAGEN GROUP

- Task
- HISTeria
- **Results**
- Contributions

HIPE 2022 HISTeria

Table 11
Final results on ajmc development dataset for all languages using best models after multi-stage fine-tuning. Results are reported with official HIPE scorer.

Submission ID	Hyperparameter Configuration	Strict F1-Score	Fuzzy F1-Score
German ($hmBERT_{32k}$) - 1	bs8-e05-lr3e-05	91.5	**94.2**
German ($hmBERT_{64k}$) - 2	bs8-e10-lr3e-05	**92.0**	93.9
English ($hmBERT_{32k}$) - 1	bs4-e10-lr3e-05	**89.1**	92.9
English ($hmBERT_{64k}$) - 2	bs8-e10-lr3e-05	88.0	**93.8**
French ($hmBERT_{32k}$) - 1	bs4-e10-lr3e-05	**86.8**	**93.1**
French ($hmBERT_{64k}$) - 2	bs4-e10-lr5e-05	85.9	93.0

HIPE 2022 HISTeria

Table 12
Final results on ajmc test dataset for all languages compared to other participants in the HIPE-2022 shared task. HISTeria denotes our system. Rank is ordered by strict F1-score.

Rank	Language	Submission ID	Strict F1-Score	Fuzzy F1-Score
1	German	L3i (team 2) - 2	93.4	95.2
2	German	HISTeria ($hmBERT_{32k}$) - 1	91.3	93.7
3	German	HISTeria ($hmBERT_{64k}$) - 2	91.2	94.5
4	German	L3i (team 2) - 1	90.8	93.4
5	German	Neural baseline	81.8	87.3
1	English	HISTeria ($hmBERT_{64k}$) - 2	85.4	91.0
2	English	L3i (team 2) - 1	85.0	89.4
3	English	L3i (team 2) - 2	84.1	88.4
4	English	HISTeria ($hmBERT_{32k}$) - 1	81.9	89.9
5	English	Neural baseline	73.6	82.8
1	French	HISTeria ($hmBERT_{64k}$) - 2	84.2	88.0
2	French	HISTeria ($hmBERT_{32k}$) - 1	83.3	88.8
3	French	L3i (team 2) - 2	82.6	87.2
4	French	L3i (team 2) - 1	79.8	86.0
5	French	Neural baseline	74.1	82.5

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

universität wien

Software Innovation Center
DATA LAB
VOLKSWAGEN GROUP

- Task
- HISTeria
- Results
- **Contributions**

HIPE 2022 HISTeria

Table 14
Community contributions for our HIPE-2022 submission: Pretrained language models and fine-tuned NER models are publicly available on the Hugging Face Model Hub.

Model Description	Model Name
$hmBERT_{32k}$ Tiny Model	dbmdz/bert-tiny-historic-multilingual-cased
$hmBERT_{32k}$ Mini Model	dbmdz/bert-mini-historic-multilingual-cased
$hmBERT_{32k}$ Small Model	dbmdz/bert-small-historic-multilingual-cased
$hmBERT_{32k}$ Medium Model	dbmdz/bert-medium-historic-multilingual-cased
$hmBERT_{32k}$ Base Model	dbmdz/bert-base-historic-multilingual-cased
$hmBERT_{64k}$ Base Model	dbmdz/bert-base-historic-multilingual-64k-td-cased
NER First Stage ($hmBERT_{32k}$)	dbmdz/flair-hipe-2022-ajmc-all
NER First Stage ($hmBERT_{64k}$)	dbmdz/flair-hipe-2022-ajmc-all-64k
NER Second Stage - German ($hmBERT_{32k}$)	dbmdz/flair-hipe-2022-ajmc-de
NER Second Stage - English ($hmBERT_{32k}$)	dbmdz/flair-hipe-2022-ajmc-en
NER Second Stage - French ($hmBERT_{32k}$)	dbmdz/flair-hipe-2022-ajmc-fr
NER Second Stage - German ($hmBERT_{64k}$)	dbmdz/flair-hipe-2022-ajmc-de-64k
NER Second Stage - English ($hmBERT_{64k}$)	dbmdz/flair-hipe-2022-ajmc-en-64k
NER Second Stage - French ($hmBERT_{64k}$)	dbmdz/flair-hipe-2022-ajmc-fr-64k

HIPE 2022 HISTeria

Contact

Luisa März

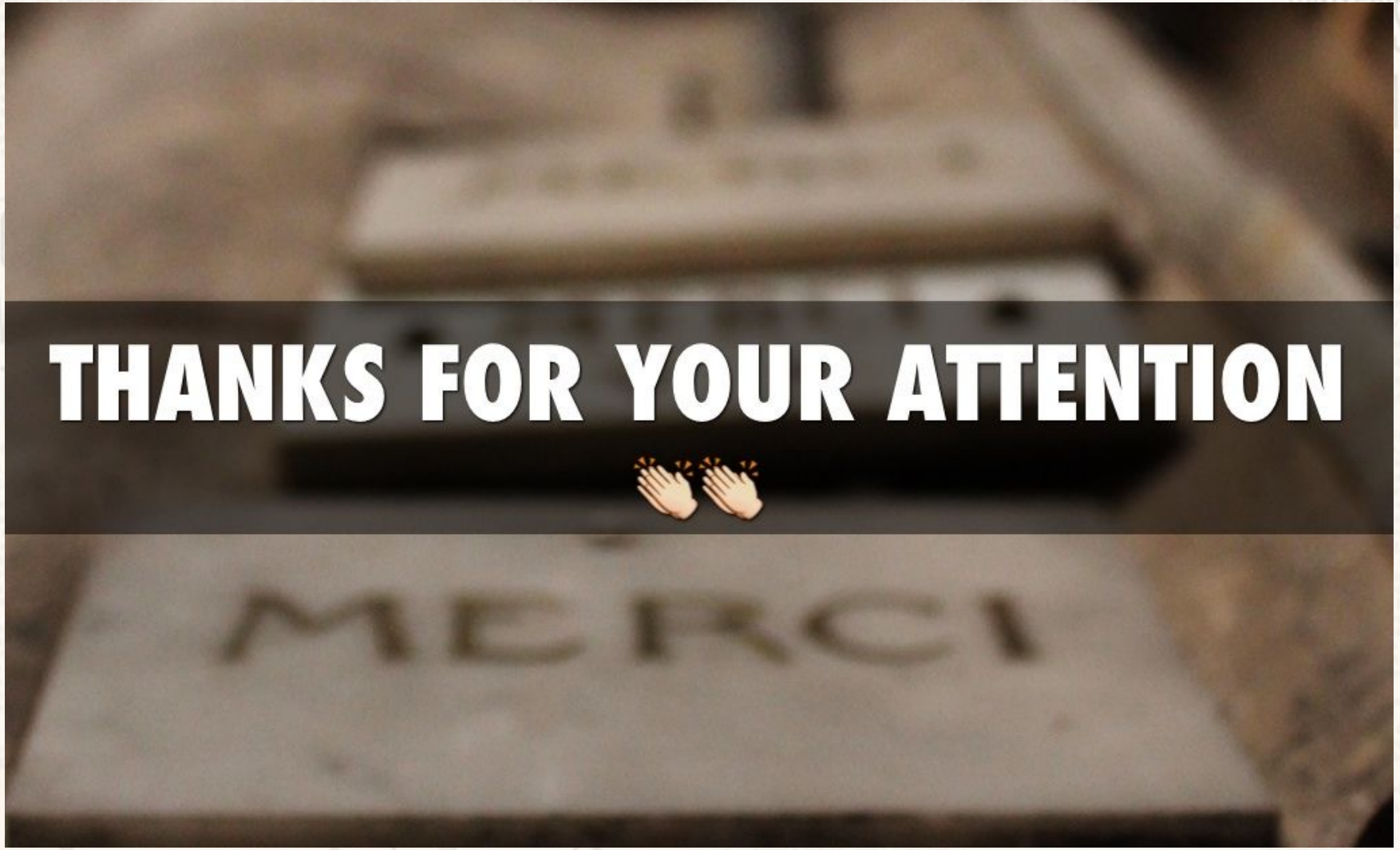
luisa.maerz@volkswagen.de
https://github.com/LuisaMaerz

Stefan Schweter

stefan@schweter.eu
https://github.com/stefan-it

Shared Task Paper:
http://ceur-ws.org/Vol-3180/paper-87.pdf

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

universität wien

Software Innovation Center
DATA LAB
VOLKSWAGEN GROUP