



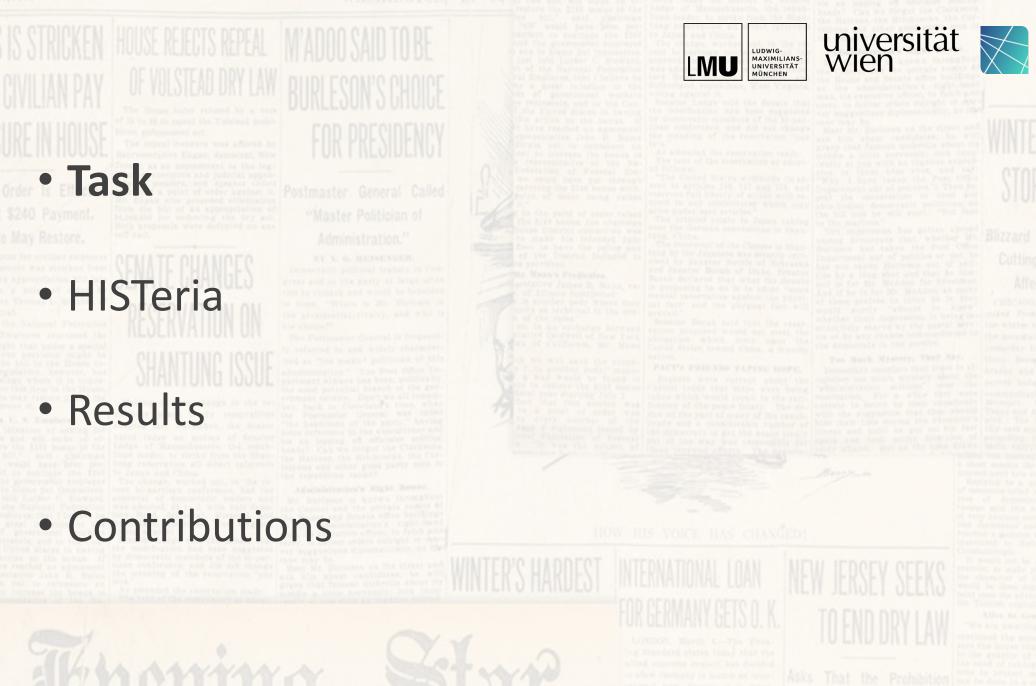


# hmBERT: Historical Multilingual Language Models for Named Entity Recognition

Stefan Schweter, Luisa März, Katharina Schmid and Erion Çano

Team HISTeria

CLEF 2022, HIPE Shared Task, 08.09.2022





Blizzard Sweeps Eastwa Cutting Communication,

Affecting Industry.

ALLIES AND AGAINST

OF THE L

here the subject ever, that part of the remference of politics in conference had in the host own contribution of the contribut

W. R. E. FARE PI







### HIPE 2022 Task

### **Multilingual Classical Commentary Challenge**

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

09.09.22

Luisa März - HIPE 2022

3





Blizzard Sweeps Eastwa Cutting Communication,

Affecting Industry.

ALLIES AN AGAINST

operating the master.

A. Promiser Libert of The State Department of the subject interaction at the similar constraint of the constraint o

W. R. E. FARE PL







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





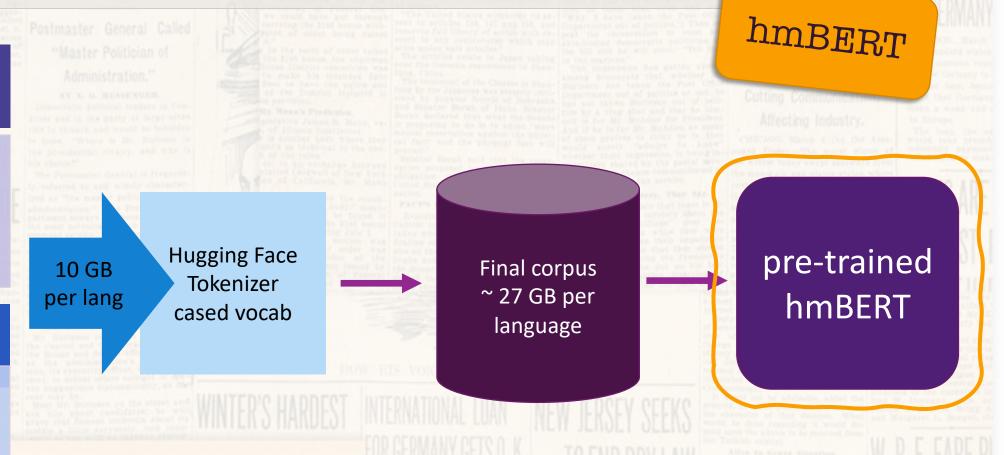


#### Europeana Newspapers

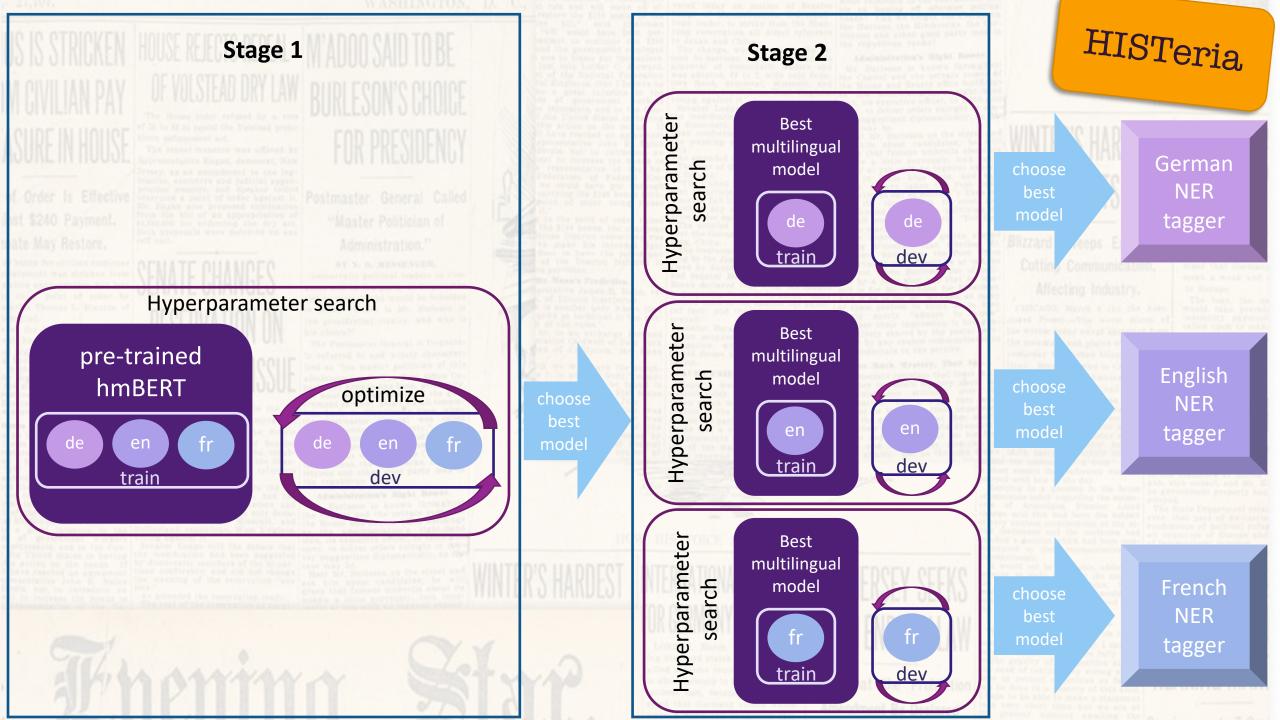
- de, fr, sv, fi
- filtered for OCR confidence

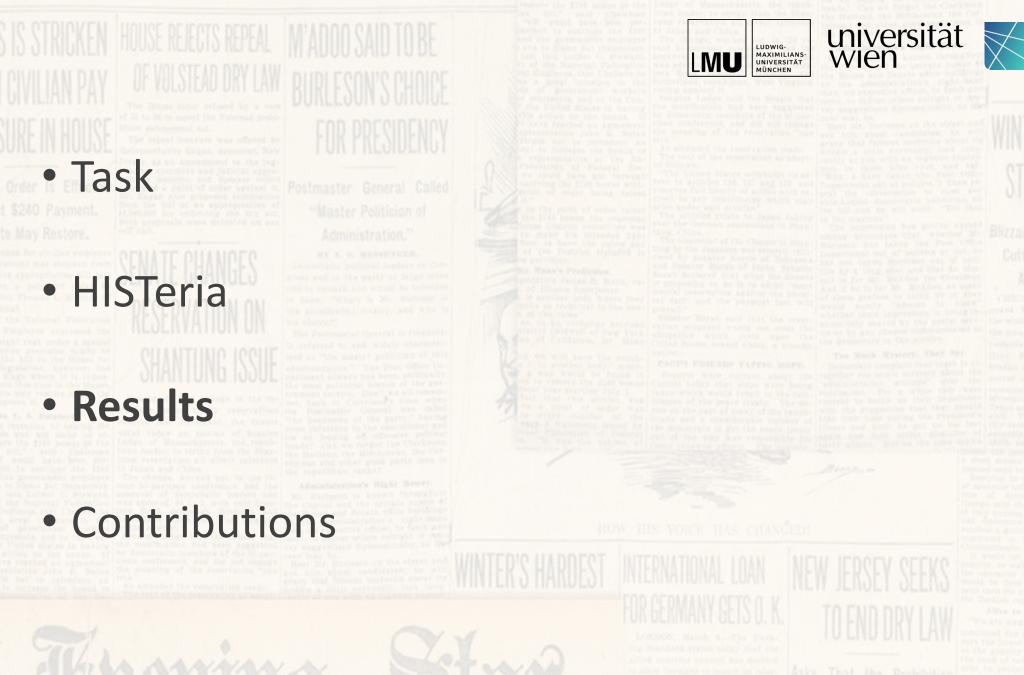
#### **Bristish Library**

- en
- english textsbetween 1800– 1900 only



Luisa März - HIPE 2022







Cutting Communication,

Affecting Industry.

at the training of the trainin

arriving addition to be a seen of the constraint of the constraint was an in the house of the constraint in the house the house the constraint in the house the hous

in outlesses had all venneries of Europe of International Integral Integral Europe of Management Integral Europe of International International Integral Integr

W. R. E. FARE PL







Table 11

Final results on a jmc 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	<b>94.2</b> 93.9
German $(hmBERT_{64k})$ - 2	bs8-e10-lr3e-05	<b>92.0</b>	
English $(hmBERT_{32k})$ - 1	bs4-e10-lr3e-05	<b>89.1</b>	92.9
English $(hmBERT_{64k})$ - 2	bs8-e10-lr3e-05	88.0	<b>93.8</b>
French $(hmBERT_{32k})$ - 1	bs4-e10-lr3e-05	<b>86.8</b>	<b>93.1</b> 93.0
French $(hmBERT_{64k})$ - 2	bs4-e10-lr5e-05	85.9	

09.09.22 Luisa März - HIPE 2022



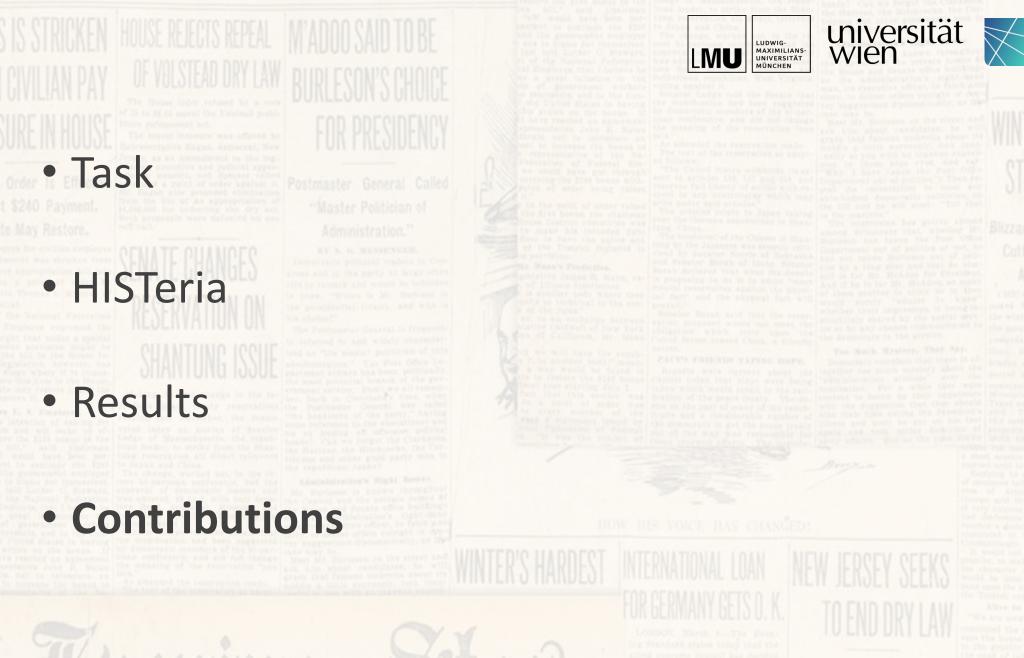




**Table 12**Final results on a jmc test dataset for all languages compared to other participants in the HIPE-2022 shared task. HISTeria denotes our system. Rank is ordered by strict F1-score.

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

09.09.22





Cutting Communication

Affecting Industry.

nd Shouth Indicate were reshored, and real and wire
father in Hanson, Obligheres,
Sorra, Arthunda, morth
for Arthunday Wiley or Support State
of Arthunday State Support State
and morth indicate and innegate.

minus. Preparating the masses of the State Department of the state of

W. R. E. FARE PL



**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 $hmBERT_{32k}$ Mini Model $hmBERT_{32k}$ Small Model $hmBERT_{32k}$ Medium Model $hmBERT_{32k}$ Base Model $hmBERT_{64k}$ Base Model	dbmdz/bert-tiny-historic-multilingual-cased dbmdz/bert-mini-historic-multilingual-cased dbmdz/bert-small-historic-multilingual-cased dbmdz/bert-medium-historic-multilingual-cased dbmdz/bert-base-historic-multilingual-cased 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

09.09.22 Luisa März - HIPE 2022







#### **Contact**

Luisa März

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

**Shared Task Paper:** 

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



**Stefan Schweter** 

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

M CIVILIAN PAY

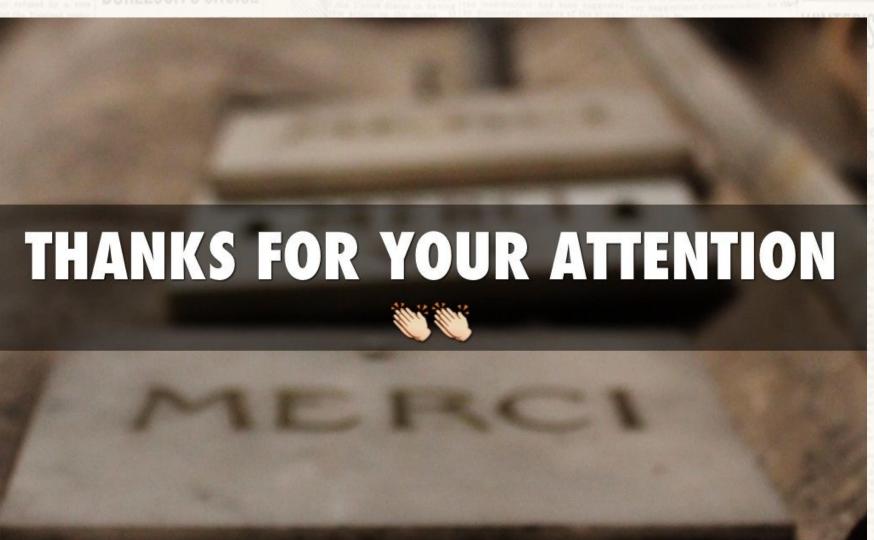
OF VOLSTEAD DRY LAW

BURLESON'S CHOIL



universität wien





EST INTERNATION

LONDON, Marci beg Standard stat affied supreme or to allow Cormany t national hom, be-

sized that derman near a weak and in Europe. The loss, the a would take proper administ payments

ALLIES ARE

AGAINST I

residence of Affred or commut, and Mr. I comment property to tale Department rehat part of Switzer cour of political refu

ational integrat.
Control terroral Kato the contulate in
Decrepts, vice on
L Deficia, Brigg A
garet L. Sunger, et

W. R. E. FARE P

HEARING MARI