SocialSec 2023
9th International Symposium on Security and Privacy in Social Networks and Big Data
University of Kent, Canterbury (UK) | August 14-16, 2023

Detection of Privacy-Harming Social Media Posts in Italian

FEDERICO PEIRETTI, RUGGERO G. PENSA

UNIVERSITY OF TURIN (ITALY)

Privacy-sensitive content

[Battaglia et al. (2020)]

User-generated content that contains sensitive information, concerning the private life of **the author or any other identifiable person** and its **explicit** or **implicit** disclosure, could potentially **cause harm or embarrassment** to them.

Content sensitivity analysis [Battaglia et al. (2020)]

Data mining task aimed at **recognizing whether a user-generated content is privacy-sensitive** or not.

Definition: Given a user-generated content item $c_i \in C$, where C is user-generated content domain, content sensitivity analysis is a task aimed at defining a function

$f_s : C \rightarrow \{sensitive, non-sensitive\}$, such that

$$f(c_i) = \begin{cases} sensitive & if\ c_i\ is\ privacy\text{-}sensitive \\ non\text{-}sensitive & otherwise. \end{cases}$$

BINARY CLASSIFICATION

CSA vs Self-disclosure

Content sensitivity analysis is a more general problem than Self-disclosure.

SELF-DISCLOSURE

Act of revealing personal information **about themselves** to others.

[Jourard S.M. (1971)]

CONTENT SENSITIVITY ANALYSIS

Contents with sensitive information regarding the private life of **the author or any other identifiable person** […]

[Battaglia et al. (2020)]

Motivation

Most of the studies focus on English texts, but…

People tend to disclose mostly in their own native languages compared to English!

[Tang et al. (2011)]

What is the problem?

Most of national/regional idioms are

LOW RESOURCE LANGUAGES

(due to the lack of large linguistic corpora to train LMs)

Our contribution

We solve CSA task for Italian using two approaches:

1. Monolingual End-to-End Learning
2. Cross-lingual Transfer Learning

First Italian corpus specifically annotated for CSA task:

ITA-SENS

Text corpora for Content Sensitivity Analysis

ITA-SENS dataset

https://github.com/federicopeiretti/ITA-SENS

We propose a new dataset ITA-SENS.

More than 15k text posts in Italian from:

TWITTER

- **FEEL-IT** [Bianchi et al. (2021)]
- **SENTIPOLC** [Barbieri et al. (2016)]
- **Italian public pages** using Twitter APIs

(posts downloaded before the new restrictions)

INSEGRETO

(«secretly» in ENG)

[https://insegreto.com/it]

- anonymous sharing of lives, secrets, opinions on different topics (health, politics, religions, love, sex…)

ITA-SENS dataset: Annotation

ANONYMITY ASSUMPTION

SENSITIVE	if the author publish it anonymously, hiding their real identity
NON SENSITIVE	if the author can be identified from it

[Correa et al. (2015)]

ITA-SENS dataset: Annotation

$\sim 8k$ posts from **Insegreto** are labeled as ***sensitive***

- published in anonymous way
- contain sensitive information
- deal with sensitive topics (politics, religion, sex…)

$\sim 7k$ posts from **Twitter** are labeled as ***non-sensitive***

- public pages and profiles

ITA-SENS dataset: Annotation

Anonymity assumption is simplistic [Bioglio et al. (2022)]

but…

Our main goal is to study whether multilingual text analysis approaches can compete with monolingual ones for CSA task.

SENS2+OMC dataset

We use an additional dataset of social media posts written in English.

Goal: solve CSA task by **transferring knowledge from English to Italian**.

SENS2

[Bioglio et al. (2022)]

- $\sim 8k$ text posts from Facebook covering a wide range of topics
- **manually annotated** according to *privacy-sensitive content* definition.
  - ✓ ***sensitive*** if post is privacy-sensitive
  - ✓ ***non-sensitive*** otherwise

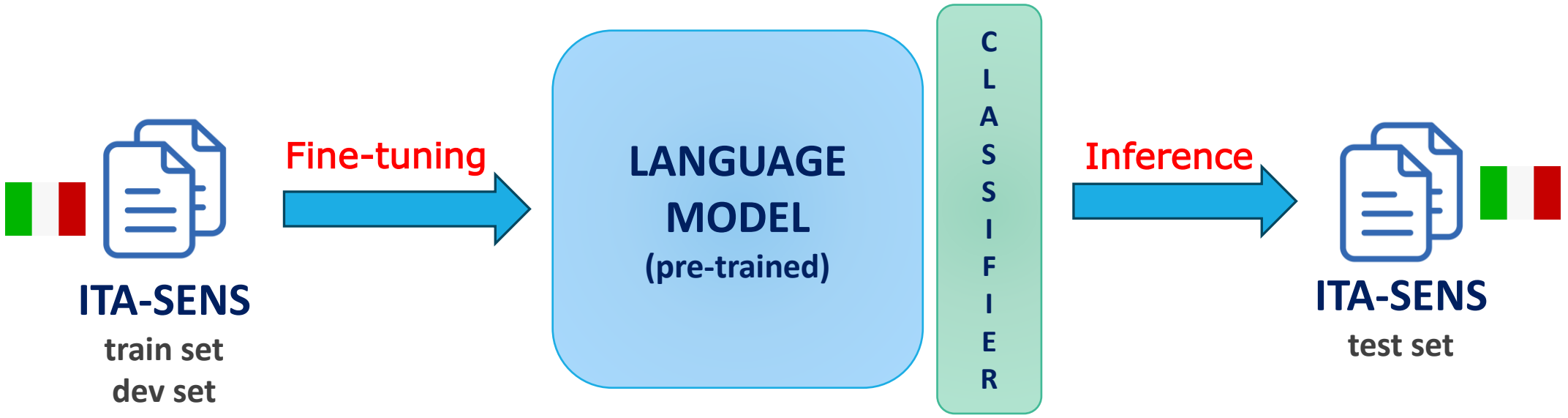
OMC

[Jaidka et al. (2020)]

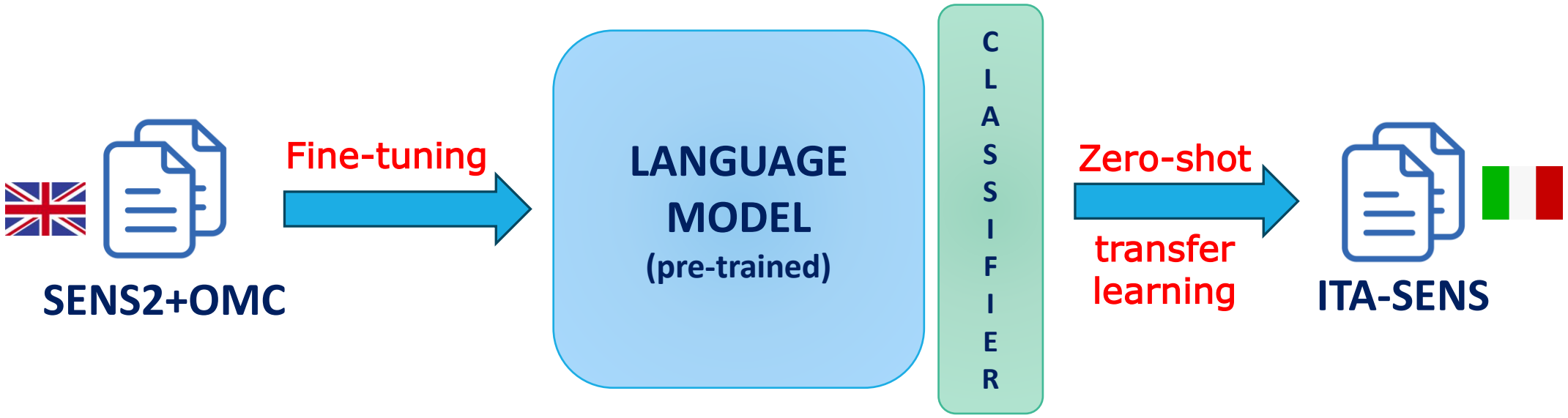
- $\sim 17k$ text posts from Reddit covering a wide range of topics
- **manually annotated** according to informational and emotional disclosure
  - ✓ ***sensitive*** if post discloses informational or emotional data
  - ✓ ***non-sensitive*** otherwise

Experimental settings

Experiment 1: Monolingual content classification



Experiment 2:
Zero-shot cross-lingual transfer learning



Pre-trained LMs

Multilingual models

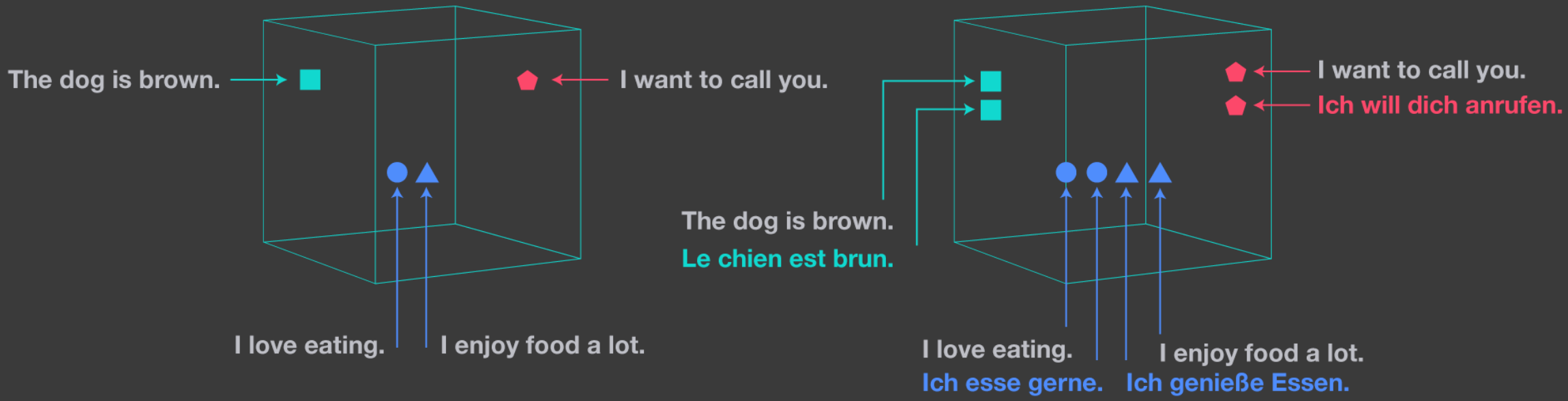
- **Multilingual BERT**
  [Devlin et al. 2018, Google]
- **XLM-RoBERTa**
  [Conneau et al. 2019, Meta AI]

Monolingual models for Italian

- **AlBERTo** based on BERT
  [Polignano et al. 2019]
- **UmBERTo** based on RoBERTa
  [Parisi et al. 2020]
- **GilBERTo** based on RoBERTa
  [Ravasio et al. 2020]

Transformer-based LMs

- HuggingFace Transformers APIs
- Linear layer for classification on top of pooled output
- AdamW optimizer with linear scheduler
- Binary Cross Entropy



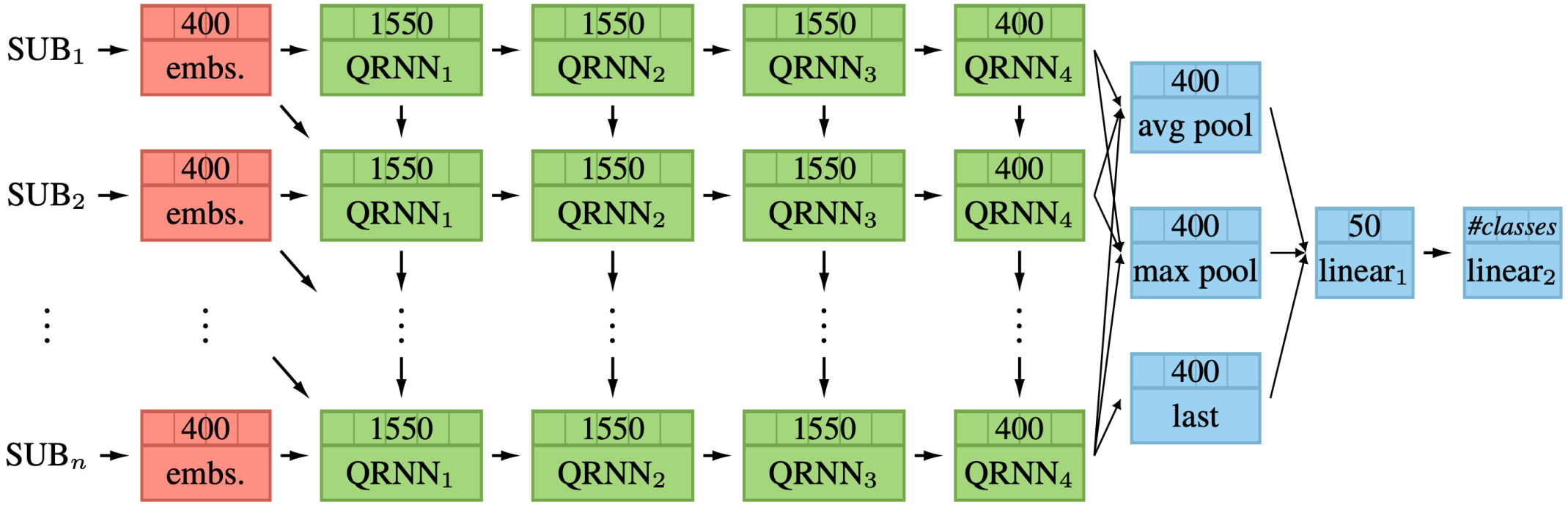
The image on the left shows a monolingual embedding space. The one on the right illustrates LASER's approach, which embeds all languages in a single, shared space. Image from [Artexte et al. (2019)]

LASER

Language-Agnostic Sentence Representations

[Artetxe et al. 2019]

- Learn joint multilingual sentence embeddings
- Encoder: Bidirectional LSTM with shared BPE vocab
- Linear layer for classification on top of the encoder
- AdamW optimizer
- Binary Cross Entropy



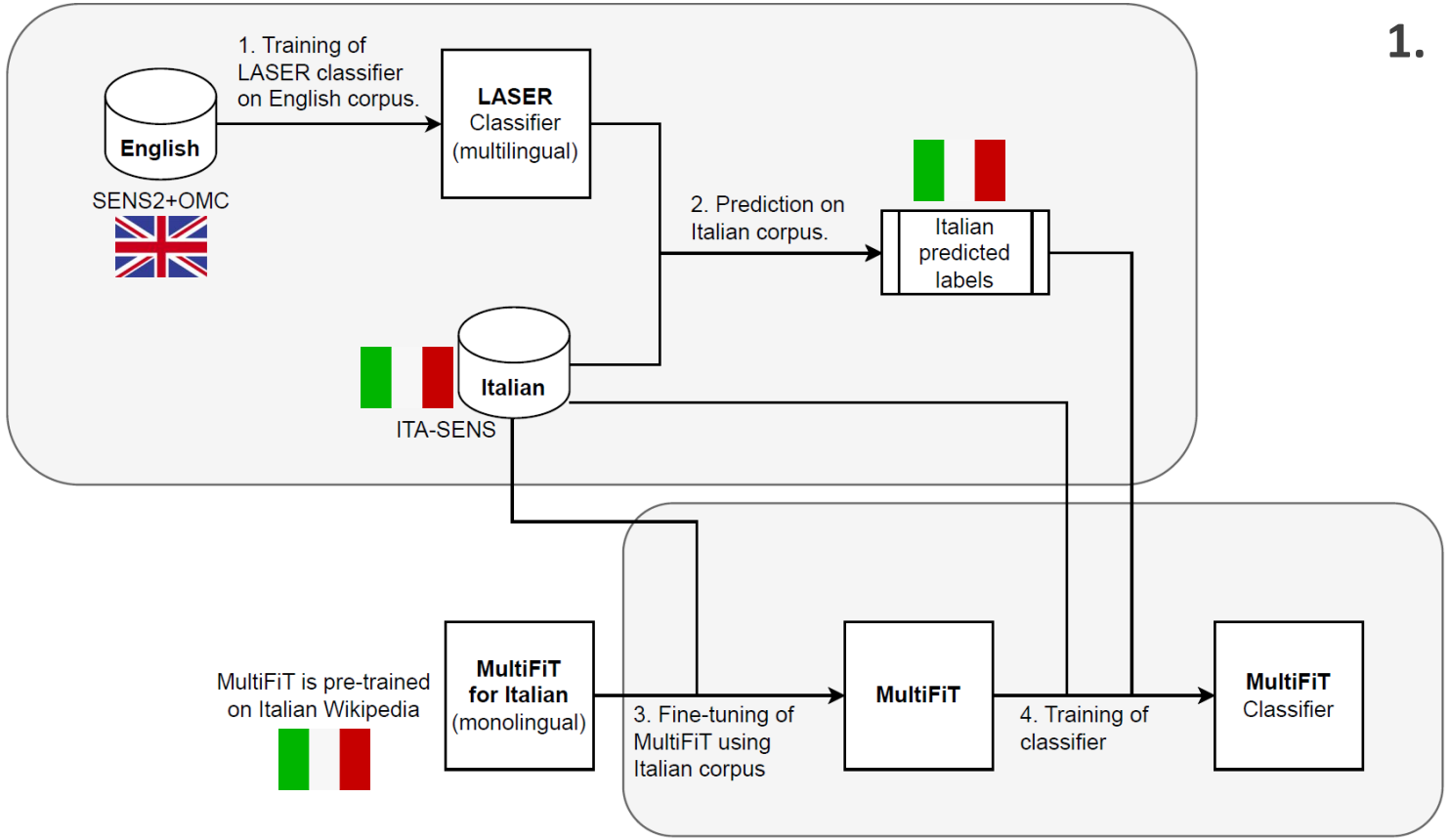
The MultiFiT language model with classifier consisting of a subword embedding layer, four QRNN layers, an aggregation layer, and two linear layers. Image from [Ruder et al. (2019)]

MultiFiT

Efficient Multi-lingual LM Fine-tuning

[Ruder et al. (2019)]

- Pre-trained on Wikipedia
- Discriminative fine-tuning (tune each layer with different learning rate)
- Robustness to label noise
- Label smoothing (overfitting, overconfidence)
- One-cycle policy with cosine annealing (scheduler)



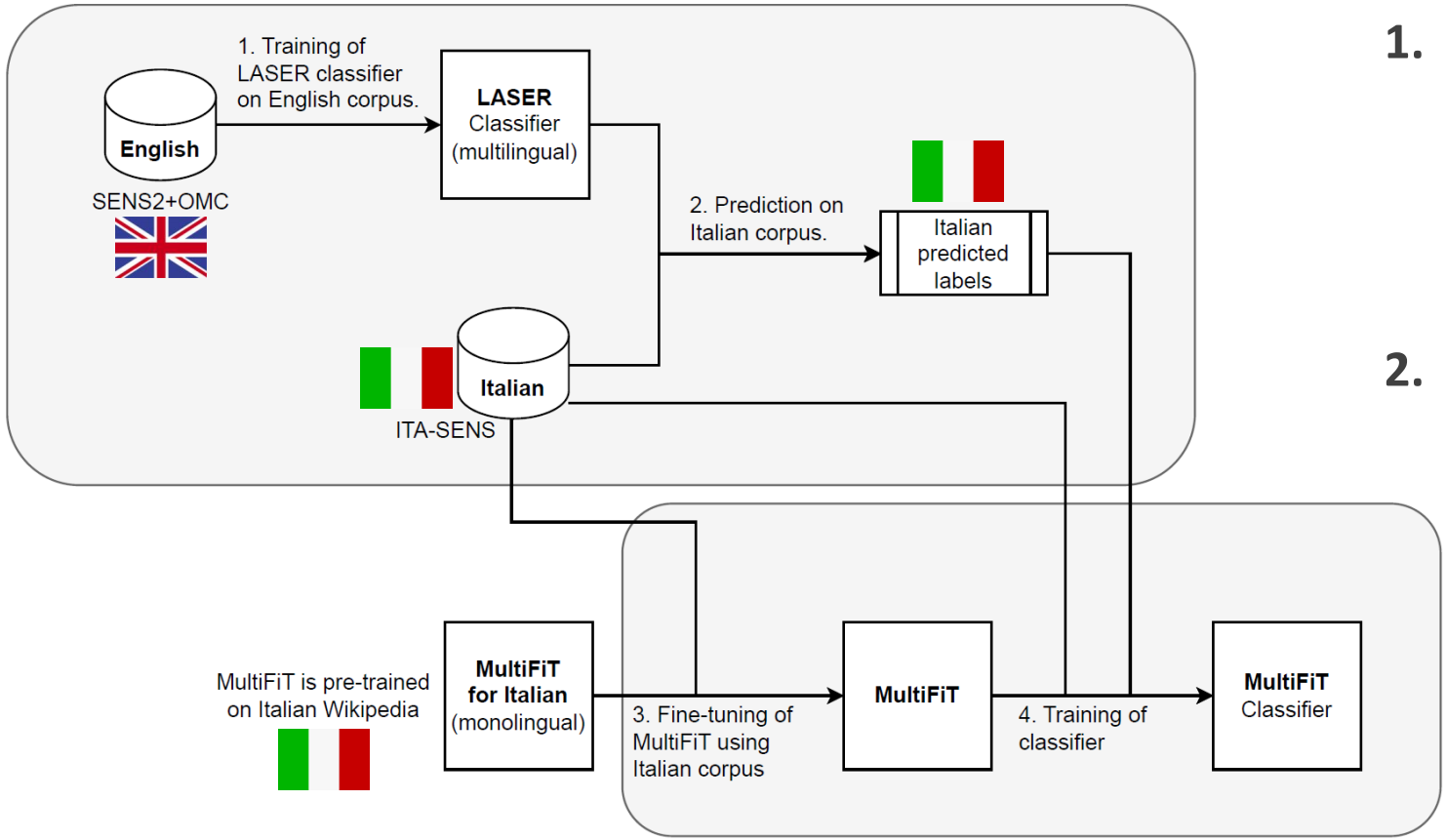
1. **Cross-lingual teacher:** LASER classifier
   - trained on SENS3+OMC (eng)
   - inference on ITA-SENS (ita)
   => PSEUDO-LABELS

MultiFiT

Zero-shot cross-lingual transfer [Ruder et al. (2019)]

Bootstrapping method adopted to perform zero-shot transfer with MultiFiT using LASER classifier as cross-lingual teacher.

- Label smoothing (avoid overfitting and overconfidence)
- Robustness to label noise



1. **Cross-lingual teacher:** LASER classifier
   - trained on SENS3+OMC (eng)
   - inference on ITA-SENS (ita)
   => PSEUDO-LABELS

2. **Zero-shot transfer** with MultiFiT for Italian
   - pre-trained on Wikipedia ITA
   - fine-tuning on ITA-SENS
   - train the classifier on top using the pseudo-labels predicted by LASER

MultiFiT

Zero-shot cross-lingual transfer [Ruder et al. (2019)]

Bootstrapping method adopted to perform zero-shot transfer with MultiFiT using LASER classifier as cross-lingual teacher.

- Label smoothing (avoid overfitting and overconfidence)
- Robustness to label noise

Fine-tuning

Hyperparameters

- 55% train set
- 25% dev set
- 20% test set

- Batch size
- Learning rate
  - Grid search over a set of pre-defined values
- #Epochs
  - Early stopping criterion on validation loss

Experimental setting	Language Model	Batch size	Learning rate	# Epochs
Traditional ITA → ITA	mBERT	32	$5 \cdot 10^{-7}$	4
	XLM-RoBERTa	32	$1 \cdot 10^{-6}$	9
	AlBERTo	32	$5 \cdot 10^{-7}$	5
	GilBERTo	16	$5 \cdot 10^{-7}$	4
	UmBERTo-wiki	32	$5 \cdot 10^{-6}$	3
	UmBERTo-commoncrawl	32	$2 \cdot 10^{-6}$	3
	LASER	32	$2 \cdot 10^{-5}$	28
	MultiFiT	20	$1 \cdot 10^{-3}$	8
Zero-shot ENG → ITA	mBERT	32	$1 \cdot 10^{-6}$	3
	XLM-RoBERTa	32	$5 \cdot 10^{-6}$	2
	LASER	32	$2 \cdot 10^{-5}$	28
	MultiFiT	20	$1 \cdot 10^{-3}$	7

Results

Evaluation metrics

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**

- **Matthews correlation coefficient (MCC)**

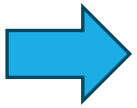
$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

- MCC is a more reliable statistical rate which produces a high score only if the prediction obtained good results in **all four confusion matrix categories (TP, TN, FP, FN)**, proportionally to the number of positive and negative samples in the dataset.
- **The greater their correlations, the more accurate the model.** [Chicco et al. (2020)]

Experiment 1:
LMs trained and tested on ITA-SENS

Model	Accuracy	Precision	Recall	F1-score	MCC
mBERT	0.8752	0.8691	0.9151	0.8722	0.746
XLM-RoBERTa	0.9181	0.9171	0.9387	0.9166	**0.927**
AlBERTo (it)	0.9260	0.9330	0.9352	0.9249	0.855
GilBERTo (it)	0.9098	0.9228	0.9157	0.9086	0.817
UmBERTo-wiki-U (it)	0.9260	0.9240	**0.9458**	0.9246	0.850
UmBERTo-CC-C (it)	0.9187	0.9347	0.9193	0.9177	0.836
LASER	0.9125	0.9168	0.9281	0.9110	0.822
MultiFiT (it)	**0.9366**	**0.9508**	0.9352	**0.9358**	0.872

Experiment 2:
Zero-shot cross-lingual transfer learning

SENS2+OMC    **ITA-SENS**

Model	Accuracy	Precision	Recall	F1-score	MCC
mBERT	0.7002	0.8458	0.5689	0.6990	0.447
XLM-RoBERTa	0.7306	**0.8814**	0.6001	0.7297	**0.508**
LASER	0.7411	0.7773	0.7544	0.7383	0.477
MultiFiT	**0.7487**	0.7879	**0.7550**	**0.7463**	0.494

Discussion and Limitations

A possible bias...

LMs may have learned to distinguish the sources of the posts (Insegreto or Twitter).

ITA-SENS is annotated following the *anonymity assumption*.

- Insegreto posts are labeled as *sensitive*
- Tweets are labeled as *non-sensitive*

An additional experiment

- We considered the classification LMs trained on **ITA-SENS**.
- Inference on a portion of $\sim 4k$ posts from the **OMC dataset**, translated into Italian using DeepL (posts with at least 20 words).
- **Baseline**: classifier assigning all posts to the majority class (sensitive)

Results

Language Model	Accuracy	Precision	Recall	F1-score	MCC
Baseline	0.7264	0.7264	1.0	0.4207	0.0
mBERT	**0.7312**	0.7519	0.9402	0.5497	0.180
XLM-RoBERTa	0.7184	0.7392	**0.9462**	0.5053	0.104
AlBERTo	0.6212	0.7999	0.6382	0.5820	0.193
GilBERTo	0.6863	0.7880	0.7771	**0.6097**	**0.220**
UmBERTo-wiki-U	0.6214	0.7741	0.6763	0.5646	0.141
UmBERTo-CC-C	0.7166	0.7380	0.9456	0.5011	0.095
LASER	0.6687	0.7857	0.7479	0.5985	0.199
MultiFiT	0.6262	**0.8103**	0.6338	0.5906	0.216

Conclusion and Future works

Conclusion

- MultiFiT outperforms all other LMs for CSA task.

- **The models trained directedly with the Italian corpus (ITA-SENS) are the best performing ones** for CSA task.

Future works

1. We plan to launch a manual annotation campaign for ITA-SENS corpus, involving several domain experts.

2. We will investigate content sensitivity analysis on images and short videos.

Thank you!

UNIVERSITÀ
DI TORINO

di.unito.it

DIPARTIMENTO
DI INFORMATICA

Questions/Comments

federico.peiretti@unito.it

ruggero.pensa@unito.it

References

Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zeroshot cross-lingual transfer and beyond. Trans. Assoc. Comput. Linguistics 7, 597–610 (2019)

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., Patti, V.: Overview of the evalita 2016 sentiment polarity classification task. In: Proceedings of CLiC-it 2016 & EVALITA 2016. CEUR-WS.org (2016)

Battaglia, E., Bioglio, L., Pensa, R.G.: Towards content sensitivity analysis. In: Berthold, M.R., Feelders, A., Krempl, G. (eds.) Proceedings of IDA 2020, Konstanz, Germany, April 27-29, 2020. pp. 67–79. Springer (2020)

Bianchi, F., Nozza, D., Hovy, D.: FEEL-IT: emotion and sentiment classification for the italian language. In: Proceedings of WASSA@EACL 2021. pp. 76–83. ACL (2021)

Bioglio, L., Pensa, R.G.: Analysis and classification of privacy-sensitive content in social media posts. EPJ Data Sci. 11(1), 12 (2022)

Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1-score and accuracy in binary classification evaluation. BMC genomics 21, 1–13 (2020)

Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 7057–7067 (2019)

References

Correa, D., Silva, L.A., Mondal, M., Benevenuto, F., Gummadi, K.P.: The Many Shades of Anonymity: Characterizing Anonymous Social Media Content. In: Proceedings of ICWSM 2015. pp. 71–80 (2015)

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

Jaidka, K., Singh, I., Liu, J., Chhaya, N., Ungar, L.: A report of the CL-Aff OffMy-Chest Shared Task: Modeling Supportiveness and Disclosure. In: Proceedings of AffCon@AAAI 2020. pp. 118–129. CEUR-WS.org (2020)

Jourard, S.M.: Self-disclosure: An experimental analysis of the transparent self. John Wiley (1971)

Parisi, L., Francia, S., Magnani, P.: Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto (2020)

Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., Basile, V., et al.: Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In: CEUR Workshop Proceedings. vol. 2481, pp. 1–6. CEUR (2019)

Ravasio, G., Di Perna, L.: Gilberto: An italian pretrained language model based on RoBERTa. https://github.com/idb-ita/GilBERTo (2020)

Ruder, S.: Neural transfer learning for natural language processing. Ph.D. thesis, NUI Galway (2019)

Tang, D., Chou, T., Drucker, N., Robertson, A., Smith, W.C., Hancock, J.T.: A tale of two languages: strategic self-disclosure via language selection on facebook. In: Proceedings of ACM CSCW 2011. pp. 387–390. ACM (2011)