Rosario Catelli

Tutor: Prof. Valentina Casola (DIETI)

co-Tutor: Eng. Concetta Pragliola (Ansaldo STS) until October 14th, 2019

co-Tutor: Dr. Massimo Esposito (ICAR CNR) from October 15th, 2019

XXXIII Cycle - III year presentation

Safeguarding privacy through deep learning techniques

My background

Graduation: M.Sc. Degree in Electronic Engineering

DIETI Group: SecLab research group

Cooperations:

- Ansaldo STS (until October 14th, 2019)
- ICAR, National Research Council (from October 15th, 2019)
- Prof. Hamido Fujita's laboratories, Iwate Prefectural University, Iwate, Japan (formally from August 1st, 2020 to November 30th, 2020)

Fellowship:

- Ansaldo STS funded Ph.D. Grant (until October 14th, 2019)
- ICAR CNR Graduate Research Grant (from October 15th, 2019)

My credits summary

Student: Rosario Catelli rosario.catelli@unina.it
Tutor: Valentina Casola valentina.casola@unina.it
Cycle XXXIII

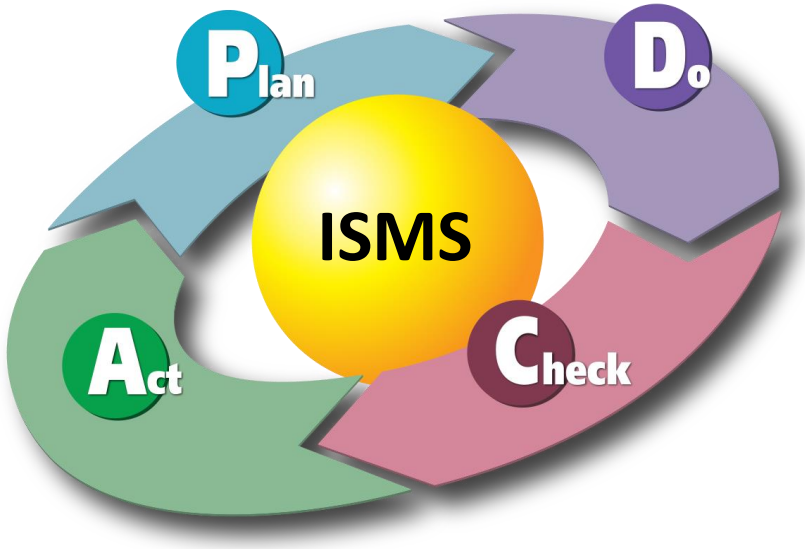
	Credits year 1								Credits year 2								Credits year 3										
		1	2	3	4	5	6			1	2	3	4	5	6			1	2	3	4	5	6	7			
	Estimated	bimonthly	bimonthly	bimonthly	bimonthly	bimonthly	bimonthly	Summary	Estimated	bimonthly	bimonthly	bimonthly	bimonthly	bimonthly	bimonthly	Summary	Estimated	bimonthly	bimonthly	bimonthly	bimonthly	bimonthly	bimonthly	bimonthly	Summary	Total	Check
Modules	20	0	0	6	3	3	7.2	**19.2**	11	5	6	2	3	0	0	**16**	0	0	0	0	0	0	0	0	**0**	**35.2**	30-70
Seminars	5	0	0	0.6	5.2	0	0	**5.8**	5	0.6	0	0.2	0	1.2	0	**2**	2.2	0	0.6	0.9	0	0	0.8	0	**2.3**	**10.1**	10-30
Research	35	10	10	3.4	1.8	7	2.8	**35**	44	4.4	4	7.8	7	8.8	10	**42**	57.8	10	9.4	9.1	10	10	9.2	10	**68**	**145**	80-140
	60	10	10	10	10	10	10	**60**	60	10	10	10	10	10	10	**60**	60	10	10	10	10	10	10	10	**70**	**190**	180

- Experience abroad
  - Prof. Hamido Fujita's laboratories at Iwate Prefectural University, Iwate, Japan (telematic collaboration due to COVID-19 pandemic, formally from August 1st, 2020 to November 30th, 2020)

My problem: general perspective

Which are the main security and privacy aspects of Information Security Management System (ISMS)?

- Compliance with security standards (e.g. ISO 27000 family)
- Compliance with privacy regulations (e.g. USA HIPAA, EU GDPR)



How and by whom are these aspects managed?

Verifying compliance, supporting auditing activities, improving risk management frameworks… all this is **solved manually** by teams of several people!

Open research challenges

PRIVACY PROBLEMS

- **Trade-off:** justifiable reasons for sharing data *VS* protect individuals' privacy
- **Legal aspect:** lack of compliance with the relevant data-protection laws
- **Social aspect:** technology "we love so much" track us indiscriminately

GENERAL PRIVACY-PROTECTION MECHANISMS

International standards propose both technical and organisational mechanisms (e.g. auditing, risk management, data integrity, …). Among them, **data minimisation and retention** focuses on so-called Personally Identifiable Information (PII), but in detail...

… HOW CAN RESEARCH HELP MINIMISE THE EXPOSURE OF PERSONAL DATA?

- Developing methods for the **de-identification** of individuals' PII captured in multimedia content (text, audio, video, …): **automated techniques!**
- Privacy By Design: **embed privacy protection systems into design specs!**

My specific research activity

Research focus

1. Electronic Health Records (EHRs)
2. Propose general de-identification methods
3. Preserve readability

- **Clinical de-identification** in medical area
- USA HIPAA defines 18 relevant identifiers, called **Protected Health Information (PHI)** (PII subgroup)
- Increasingly precise classifications: **binary**, by **category,** by **subcategory**, per **token** or per **entity**

Table 1.1: Excerpt from 45 CFR §164.514

#	PHI Identifiers
(A)	Names;
(B)	All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
(C)	All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
(D)	Telephone numbers;
(E)	Fax numbers;
(F)	Electronic mail addresses;
(G)	Social security numbers;
(H)	Medical record numbers;
(I)	Health plan beneficiary numbers;
(J)	Account numbers;
(K)	Certificate/license numbers;
(L)	Vehicle identifiers and serial numbers, including license plate numbers;
(M)	Device identifiers and serial numbers;
(N)	Web Universal Resource Locators (URLs);
(O)	Internet Protocol (IP) address numbers;
(P)	Biometric identifiers, including finger and voice prints;
(Q)	Full face photographic images and any comparable images; and
(R)	Any other unique identifying number, characteristic, or code [...]

General de-identification techniques

How to identify PHI?

Through **Named Entity Recognition** (**NER**), a Natural Language Processing (NLP) technique

State-of-the-art for clinical de-identification

Recent challenges (e.g. i2b2 and ShARe/CLEF eHealth Evaluation Lab) have driven progress:

- **Rudimentary rule-based techniques**: semantic dictionaries, gazetteers and patterns

Pros: Easy to implement

Cons: Fine-tuning necessary for each data change, non context-aware

- **Machine learning-based techniques**: classifiers and sequence labeller

Pros: "Flexible" recognition

Cons: Data-dependent, poor in complex and rare cases, time consuming feature engineering, non context-aware

- **Deep learning-based techniques**: embeddings plus neural networks

Pros: Extremely task-adaptable, potentially context-aware

Cons: Data-dependent, poor in complex and rare cases

My specific research activity: open challenges

Starting questions and problems for the specific problem

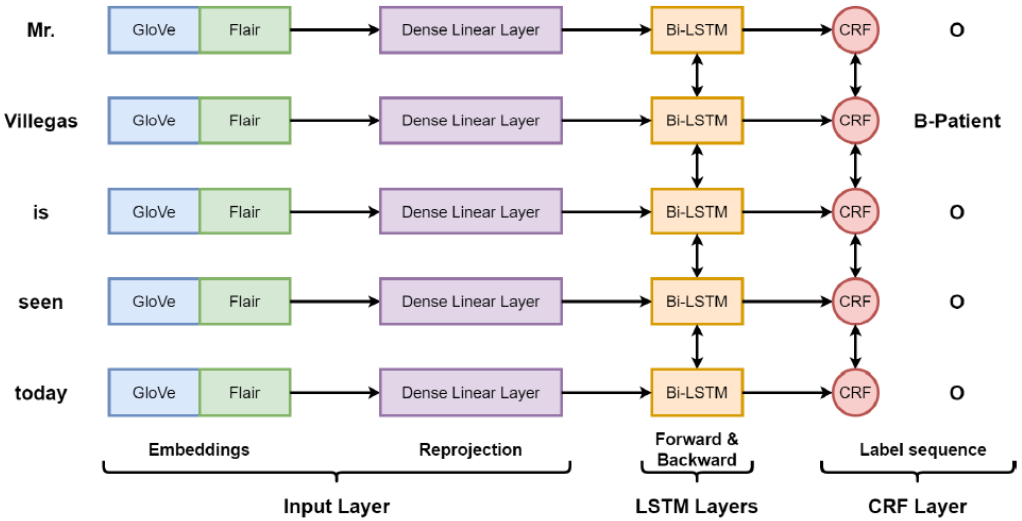
1. **Which are the best deep learning architecture and tokens representation?**
- Choose how to treat the text (e.g. words, sub-words or characters)
- Understand whether context sensitive management can deliver improvement

2. **What happens if the language domain of interest is not English?**
- EHRs are generally available in English
- Techniques are generally developed and tested for English
- No scientific literature for Italian

3. **Can the English domain be exploited to the advantage of a low-resource domain such as the Italian one?**
- Many generic multilingual alternatives
- None specific to the scenario of interest

Technique 1: a novel deep learning architecture

- **Bi-LSTM + CRF neural network** to exploit sequence modelling and labelling capabilities
- **Stacked input embeddings**: GloVe to capture latent syntactic and semantic similarities, plus Flair to interpret context and grasp morpho-syntactic variations



- **Sentences Grouping Factor (SGF)**: increase input instances to broaden context representation leveraging at its best Bi-LSTM + CRF memory capacity

Use Case 1: results

Table 4.1: Sub-document level analysis - Micro-Averaged F_1 scores

	Entity Level			Token Level		
SGF	Sub-category	Category	Binary	Sub-category	Category	Binary
1	0.9239	0.9445	0.9486	0.9443	0.9713	0.9756
2	0.9343	0.9481	0.9508	0.9581	0.9741	0.9784
4	0.9409	0.9523	0.9559	0.9631	0.9760	0.9805
8	0.9435	0.9545	0.9581	0.9646	0.9774	0.9821
16	0.9446	0.9552	0.9590	0.9655	0.9776	0.9822
32	**0.9480**	**0.9579**	**0.9614**	**0.9678**	**0.9792**	**0.9832**

Table 4.5: Micro-Averaged F_1 scores comparison

	Entity Level			Token Level		
Model	Sub-category	Category	Binary	Sub-category	Category	Binary
H. Yang and Garibaldi 2015	NA	0.9360	NA	NA	0.9611	NA
Z. Liu, Tang, et al. 2017	NA	0.9511	0.9650	NA	0.9698	0.9828
Y. Kim, Heider, and Stéphane M. Meystre 2018	NA	0.9573	NA	NA	NA	NA
Tang, D. Jiang, et al. 2019	NA	0.9550	**0.9685**	NA	0.9748	**0.9870**
$SGF = 32$	0.9480	**0.9579**	0.9614	0.9678	**0.9792**	0.9832

Use Case 1: ablation

Moreover, through ablation analysis it was possible to further confirm the validity of the proposed solution for the English language.

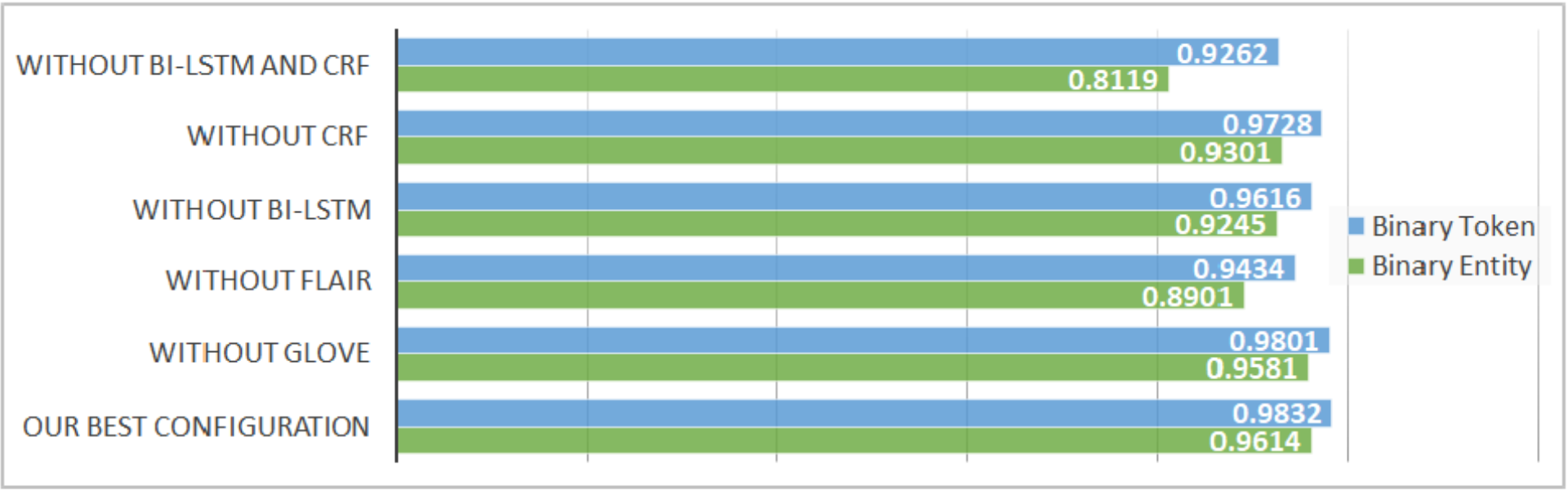


Figure 4.3: Ablation test performance.

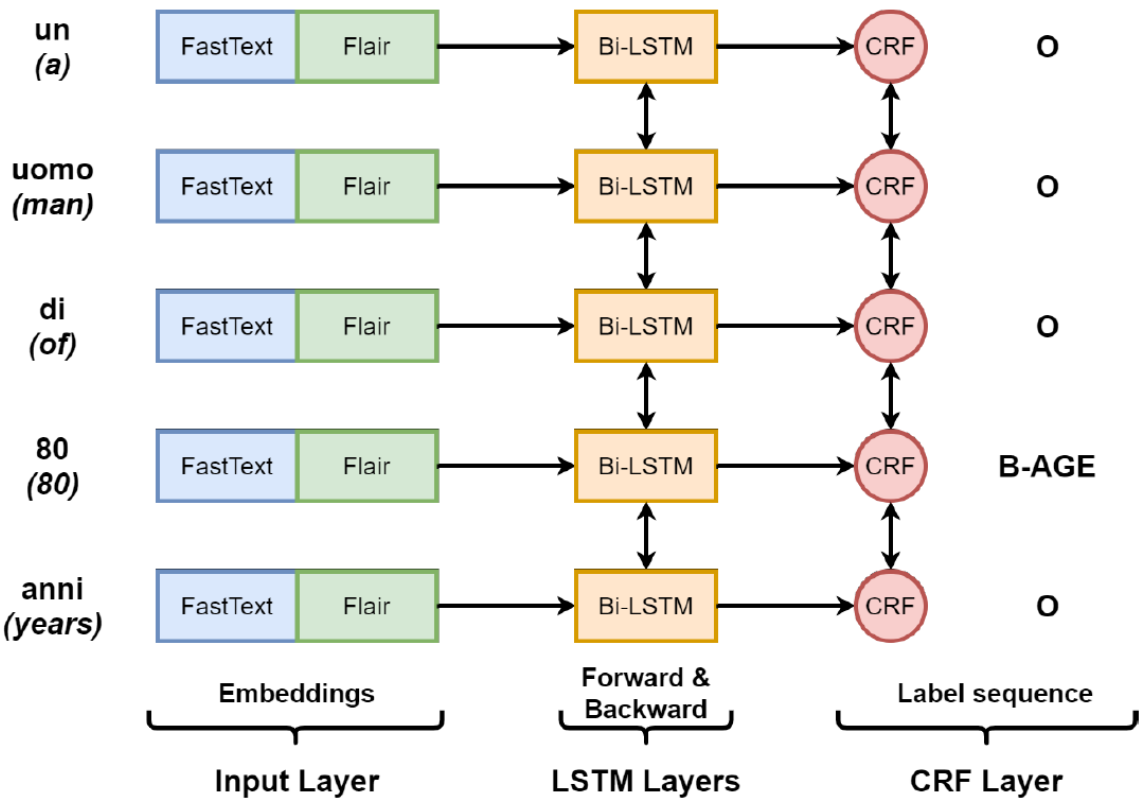
Technique 2 focused on the Italian language scenario

- **Creation of the first Italian data set for clinical de-identification**

Starting from the COVID-19 medical records made available to the public in pdf format by the Italian Society of Radiology (SIRM)

- **Build of the best deep learning architecture for Italian clinical de-identification**

Starting from the Bi-LSTM + CRF model and finding the best input layer for the specific scenario, leveraging the experience of Use Case 1



Use Case 2: results

The Italian language, similarly to the English language, presents a wide syntactic and morphological variety:

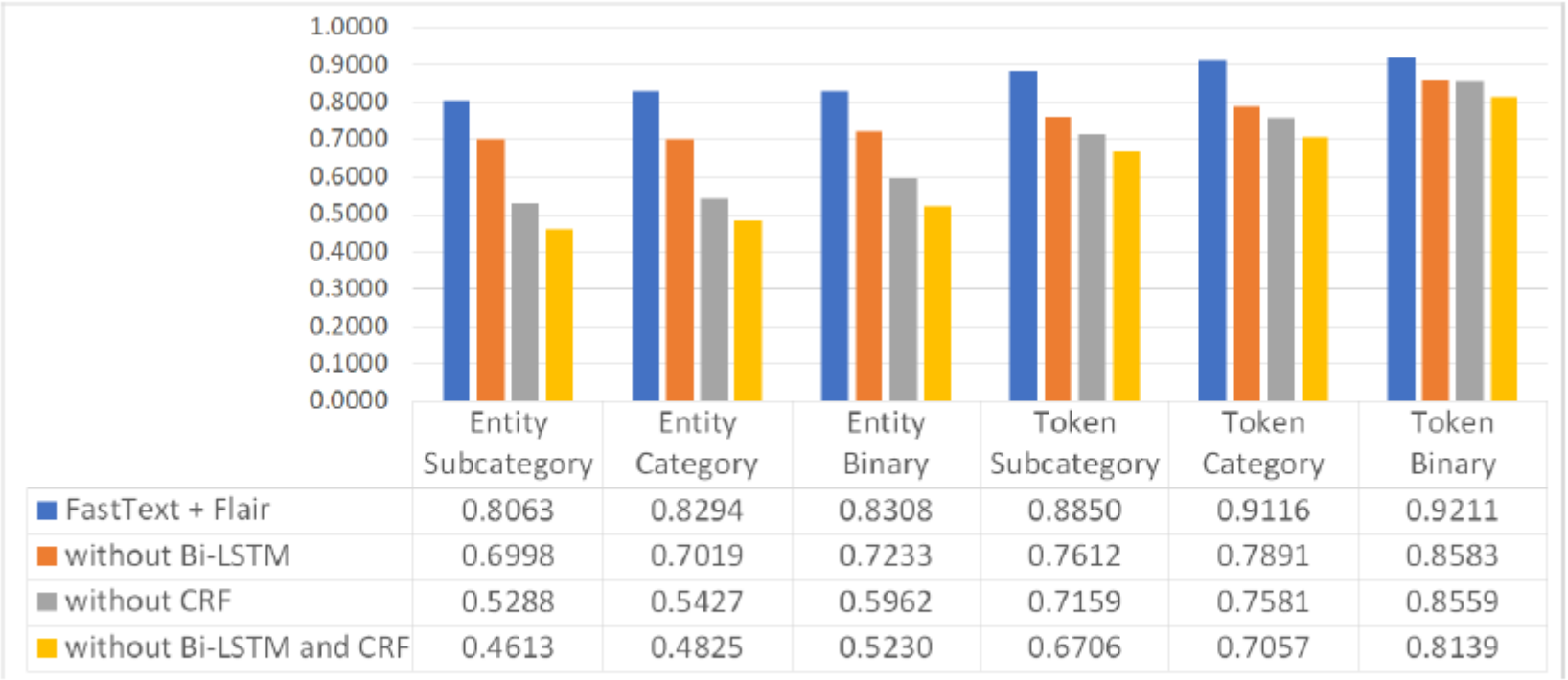
- The stacked embeddings solution for Italian outperformed the others
- BERT, a more recent NLP architectures by Google, was also outperformed

Table 4.11: Micro-Averaged F_1 results.

Model	Embedding	Entity Level			Token Level		
		Subcategory	Category	Binary	Subcategory	Category	Binary
Bi-LSTM + CRF	FastText	0.7034	0.7130	0.7297	0.7821	0.8155	0.8395
Bi-LSTM + CRF	Flair	**0.8100**	0.8224	0.8289	0.8797	0.9045	0.9211
Bi-LSTM + CRF	FastText + Flair	0.8063	**0.8294**	**0.8308**	**0.8850**	**0.9116**	0.9211
$\text{BERT}_{\text{BASE}}$ Uncased	-	0.6442	0.6667	0.6848	0.7667	0.8083	0.8796
$\text{BERT}_{\text{BASE}}$ Cased	-	0.7553	0.7880	0.7969	0.8561	0.8979	**0.9260**

Use Case 2: ablation

Also in this case, through ablation analysis, it was possible to further confirm the validity of the proposed solution for the Italian language.



	Entity Subcategory	Entity Category	Entity Binary	Token Subcategory	Token Category	Token Binary
FastText + Flair	0.8063	0.8294	0.8308	0.8850	0.9116	0.9211
without Bi-LSTM	0.6998	0.7019	0.7233	0.7612	0.7891	0.8583
without CRF	0.5288	0.5427	0.5962	0.7159	0.7581	0.8559
without Bi-LSTM and CRF	0.4613	0.4825	0.5230	0.6706	0.7057	0.8139

Figure 4.4: Ablation analysis.

Technique 3 focused on multilingual training strategies

- Investigating the ability of multilingual methods to transfer knowledge between different languages
- Find a multilingual approach specific to both de-identification and Italian language
- Extend the English dataset with the Italian one in a coherent way and understand how to exploit this condition

Use Case 3: results

Table 4.17: Micro-Averaged F_1 results

Model	Strategy	S_E	C_E	B_E	S_T	C_T	B_T
Bi-LSTM+CRF: BPEmb (IT) + Flair (IT)	IT	0.8110	0.8278	0.8317	0.8856	0.9115	0.9190
Bi-LSTM+CRF: MultiBPEmb + Flair multi fast	EN	0.2662	0.2948	0.3134	0.4103	0.4914	0.5797
	IT	0.7910	0.8118	0.8159	0.8826	0.9060	0.9183
	MIX	0.8371	**0.8602**	0.8618	0.8970	0.9304	0.9417
	EN-IT	**0.8391**	0.8595	0.8619	**0.9033**	0.9321	**0.9449**
$BERT_{base}$ (IT) Cased	IT	0.7553	0.7880	0.8561	0.7969	0.8979	0.9260
mBERT Cased	EN	0.4585	0.5029	0.6878	0.5498	0.6097	0.6878
	IT	0.7768	0.8207	**0.9449**	0.8923	**0.9353**	**0.9449**
	MIX	0.7696	0.8105	0.9379	0.8833	0.9245	0.9379
	EN-IT	0.7228	0.7576	0.8969	0.8241	0.8678	0.8969

Table 4.18: Ablation analysis

Embedding	S_E	C_E	B_E	S_T	C_T	B_T
MultiBPEmb	0.7614	0.7743	0.7914	0.8201	0.8569	0.8835
Flair multi fast	0.7621	0.7851	0.7972	0.8529	0.8801	0.8963
MultiBPEmb + Flair multi fast	**0.8391**	**0.8595**	**0.8619**	**0.9033**	**0.9321**	**0.9449**

Conclusions and possible future works

Cutting-edge results were obtained:

1. Best results for each technique tested compared to BERT and other approaches
2. Improved dependence on training data by working at character level and extending the context: morpho-syntactic variations and rare and complex data in texts are caught and better managed
3. Better training strategy for low-resource languages using multilingual scenarios

Possible future works:

1. Creation of larger and possibly multilingual de-identification data sets
2. Evaluation and comparison of other Italian versions of BERT
3. Experimenting with new data pre-processing and post-processing techniques

My products

- Journal papers (most recent accepted on top)
  - **Rosario Catelli**, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito (2021). "A Novel COVID-19 Data Set and an Effective Deep Learning Approach for the De-Identification of Italian Medical Records". In: IEEE Access 9, pp. 19097–19110. doi: 10.1109/access.2021.3054479. url: https://doi.org/10.1109/access.2021.3054479
  - Marco Pota, Mirko Ventura, **Rosario Catelli**, and Massimo Esposito (Dec. 2020). "An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian". In: Sensors 21.1, p. 133. doi: 10.3390/s21010133. url: https://doi.org/10.3390/s21010133
  - **Rosario Catelli**, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito (Feb. 2021). "Combining contextualized word representation and sub-document level analysis through Bi-LSTM+CRF architecture for clinical deidentification". In: Knowledge-Based Systems 213, p. 106649. doi: 10.1016/j.knosys.2020.106649. url: https://doi.org/10.1016/j.knosys.2020.106649
  - **Rosario Catelli**, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito (Dec. 2020). "Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set". In: Applied Soft Computing 97, p. 106779. doi: 10.1016/j.asoc.2020.106779. url: https://doi.org/10.1016%2Fj.asoc.2020.106779
  - Hassan Mokalled, **Rosario Catelli**, Valentina Casola, Daniele Debertol, Ermete Meda, and Rodolfo Zunino (2020). "The Guidelines to Adopt an Applicable SIEM Solution". In: Journal of Information Security 11.01, pp. 46–70. doi: 10.4236/jis.2020.111003. url: https://doi.org/10.4236%2Fjis.2020.111003
- Conference papers (most recent accepted on top)
  - Valentina Casola and **Rosario Catelli** (Nov. 2020). "Semantic Management of Enterprise Information Systems through Ontologies". In: Computer Science & Information Technology (CS & IT). AIRCC Publishing Corporation. doi: 10.5121/csit.2020.101403. url: https://doi.org/10.5121%2Fcsit.2020.101403
  - Valentina Casola, **Rosario Catelli**, and Alessandra De Benedictis (June 2019). "A First Step Towards an ISO-Based Information Security Domain Ontology". In: 2019 IEEE 28° International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). IEEE. doi: 10.1109/wetice.2019.00075. url: https://doi.org/10.1109%2Fwetice.2019.00075
  - Hassan Mokalled, **Rosario Catelli**, Valentina Casola, Daniele Debertol, Ermete Meda, and Rodolfo Zunino (June 2019). "The Applicability of a SIEM Solution: Requirements and Evaluation". In: 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). IEEE. doi: 10.1109/wetice.2019.00036. url: https://doi.org/10.1109%2Fwetice.2019.00036

Thank you!

Questions?

rosario.catelli@unina.it

rosario.catelli@icar.cnr.it