Analysis of Italian social media texts: from tools and resources to applications

Andrea Cimino

ItaliaNLP Lab

Institute for Computational Linguistics "Antonio Zampolli" (CNR - Pisa - Italy)

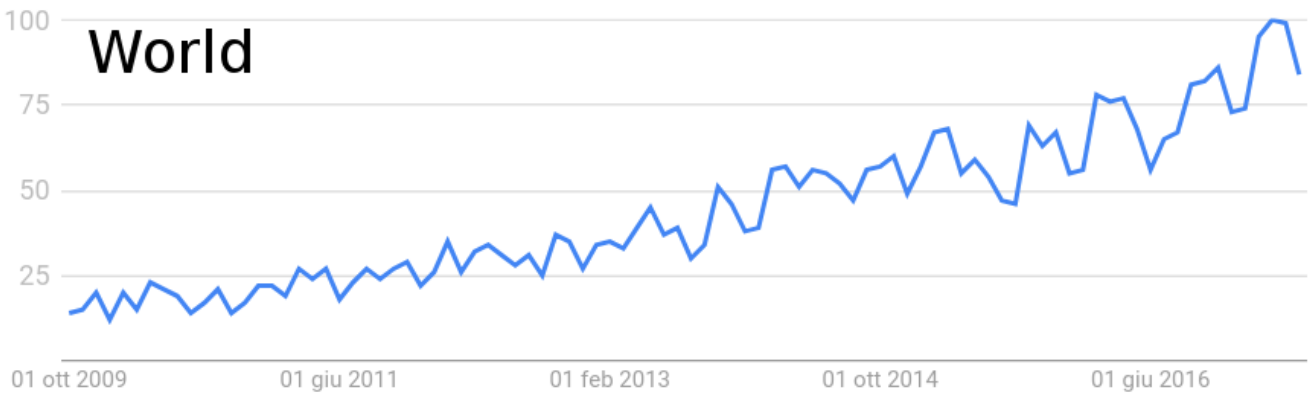
www.italianlp.it

Social media analysis

- Research in automatic extraction of data from social media analysis has flourished in the past years
  - Subjectivity Analysis
  - Sentiment Analysis
  - Irony Detection
  - ...
- This is mostly due to the rapid expansion of the Social Web (Twitter, Facebook, ...)
- But still a challenging task and a hot research topic (2016: 41.000 results on Google Scholar w.r.t. "Sentiment Analysis")

Google Trends: "Sentiment Analysis"

Interest in the analysis of social media has grown also in Italy (political communication, brand reputation, . . .)



The need of domain adaptation

	Written Language	Spoken Language	Twitter
Adjectives	6.96	4.63	3.88
Conjunctions	4.60	5.69	2.68
Interjections	0.07	1.33	0.18
Prepositions	15.30	10.50	12.38
Nouns	24.98	17.91	35.97
Verbs	13.85	17.27	9.60
Sentence length	24.21	9.18	13.16

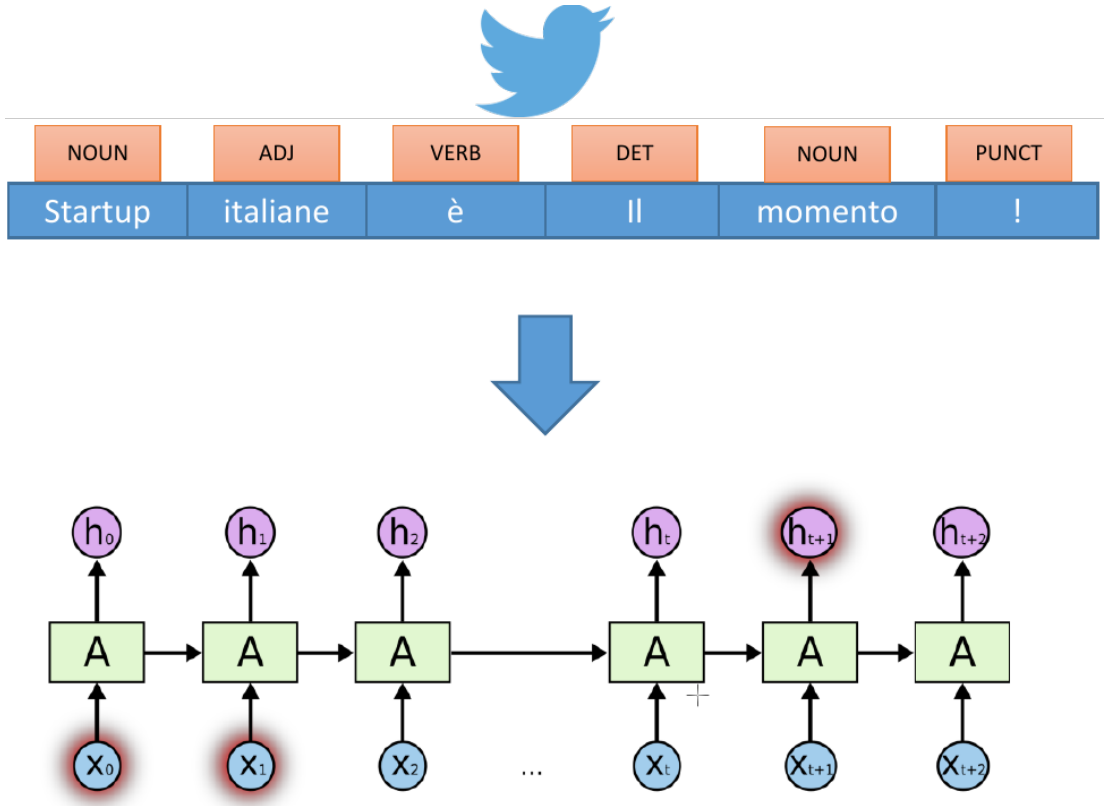
- Problem: existing NLP tools (tokenizers, pos-taggers, parsers, . . .) are trained on generic text (e.g. journals)
- Such tools have lower performances on different type of text, such as social media languages.
- NLP chains suffer from error propagation!
- Solution: annotation of corpora and development of adapted NLP tools

The Italian community effort

- Several corpora were manually annotated for different NLP tasks:
  - PoS-Tagging (6,768 Tweets)
  - Named Entity Recognition (9,410 Tweets)
  - Sentiment Analysis (9,410 Tweets)
  - Irony Detection (9,410 Tweets)
  - Subjectivity Detection (9,410 Tweets)
- Corpora were exploited by the Italian NLP community to build NLP tools
- Tools evaluated at EVALITA, the evaluation campaign of NLP tools for Italian
- Contributions from more than 10 Italian universities and research institutes

POS tagging for Italian Social Media Texts

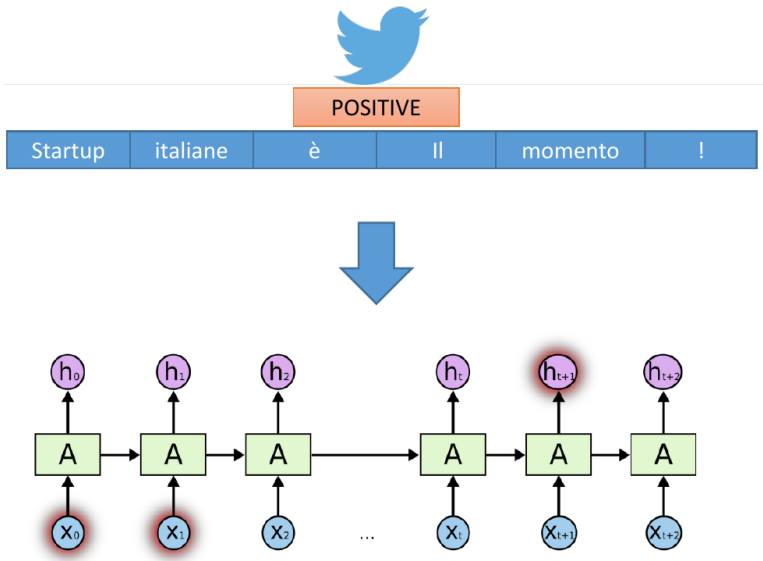
- The ItalianNLP Lab developed the best postagger
- Based on Deep Neural Networks (LSTM)
- Ranked 1st among 9 participants
- 93.2% accuracy



Andrea Cimino, Felice Dell'Orletta: *Building the state-of-the-art in POS tagging of Italian Tweets*. CLiC-it/EVALITA 2016

Sentiment Analysis

- Deep learning is the most performant learning technique
- LSTM are able to capture long relationships between words in a sentence
- Polarity changes are the most challenging to detect: *I don't think tomorrow will be a nice day*
- Twitter corpus labeled with 4 classes: positive, negative, neutral, positive-negative
- Best F-score: 0.66 by University of Pisa



Giuseppe Attardi, Daniele Sartiano, Chiara Alzetta, Federica Semplici: *Convolutional Neural Networks for Sentiment Analysis on Italian Tweets*. CLiC-it/EVALITA 2016

Social Sensing Applications

- In addition to the NLP task previously described, ongoing works to exploit social media users as **social sensors**
- Joint works between ItalianNLP Lab (ILC-CNR) and Wafi Lab (IIT-CNR)
  - **Damage assessment**: extracting real time information.
  - **Hate speech detection**: detection of flames on the web.
  - **Witness detection**: finding viewers of an event.

Damage assessment via Social Media Analysis

- People turns to social media in the aftermath of disasters to seek and publish critical and up to date information
- Automatically generated crisis maps can detect both highly and lightly damaged areas
- Goal: prioritize rescue efforts where they are most needed

Damage assessment via Social Media Analysis

- Goal: develop a prototype platform for decision support in emergency management that can analyze in real time the content shared by users
- Manually annotated tweets as: *damage*, *no damage*, *not relevant*

			Tweets			
Dataset	**Type**	**Year**	*damage*	*no damage*	*not-relevant*	*Total*
L'Aquila	Earthquake	2009	563	312	480	270
Emilia	Earthquake	2012	2,761	507	2,141	522
Sardegna	Flood	2013	597	717	194	65

Damage assessment via Social Media Analysis

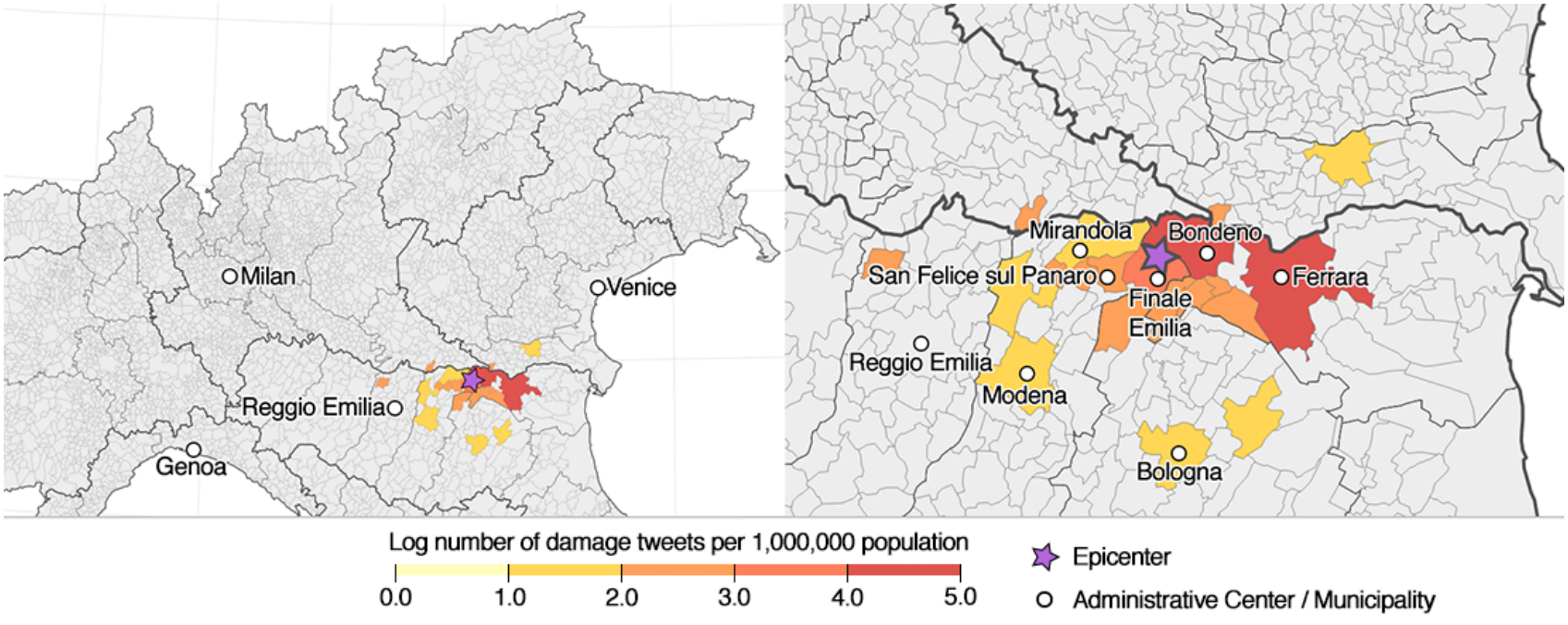
We built an SVM based classifier exploiting:

- raw and lexical text features
- morpho-syntactic features
- syntactic features
- sentiment analysis features

		damage			*no damage*			*not relevant*		
Dataset	**Accuracy**	**Prec.**	**Rec.**	**F-M.**	**Prec.**	**Rec.**	**F-M.**	**Prec.**	**Rec.**	**F-M.**
L'Aquila	0.83	0.92	0.87	0.89	0.81	0.87	0.84	0.77	0.71	0.74
Emilia	0.82	0.91	0.88	0.90	0.85	0.89	0.87	0.54	0.46	0.50
Sardegna	0.78	0.86	0.93	0.89	0.50	0.46	0.48	0.31	0.14	0.19

Results of the damage detection task on the available datasets.

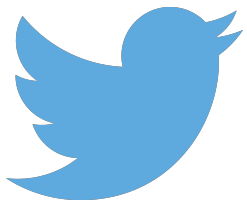
Damage assessment via Social Media Analysis



Distribution of damage tweets among the municipalities of Northern Italy (Emilia 2012)

Stefano Cresci, Maurizio Tesconi, Andrea Cimino, Felice Dell'Orletta: *A Linguistically-driven Approach to Cross-Event Damage Assessment of Natural Disasters from Social Media Messages*. WWW 2015.

Thanks for your attention!
Questions?



References

- Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, Viviana Patti: Overview of the Evalita 2016 SENTIment POLarity Classification Task. CLiC-it/EVALITA 2016
- Cristina Bosco, Fabio Tamburini, Andrea Bolioli, Alessandro Mazzei: Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian Task. CLiC-it/EVALITA 2016
- Giuseppe Attardi, Daniele Sartiano, Chiara Alzetta, Federica Semplici: Convolutional Neural Networks for Sentiment Analysis on Italian Tweets. CLiC-it/EVALITA 2016
- Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, Maurizio Tesconi: Hate Me, Hate Me Not: Hate Speech Detection on Facebook. ITASEC 2017: 86-95

References

- Andrea Cimino, Felice Dell'Orletta: Building the state-of-the-art in POS tagging of Italian Tweets. CLiC-it/EVALITA 2016
- Stefano Cresci, Andrea Cimino, Felice Dell'Orletta, Maurizio Tesconi: Crisis Mapping During Natural Disasters via Text Analysis of Social Media Messages. WISE, 2015