Not on my (Neural) Watch

Automatic Detection of Hate Speech in Social Media

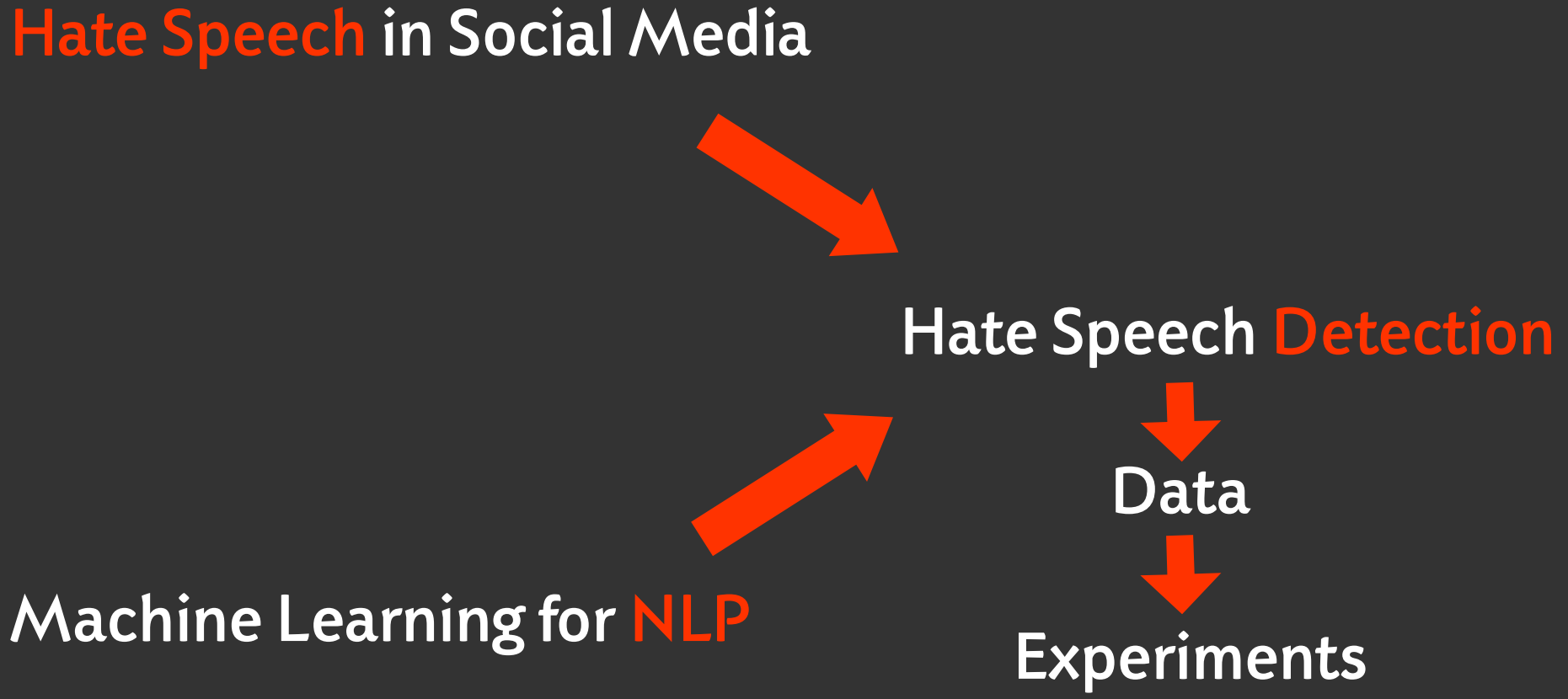
Valerio Basile
25/1/2019

> whoami

Msc Computer Science (here!)
PhD in Computational Semantics
Postdoc Research Fellow at University of Turin

http://hatespeech.di.unito.it/
http://valeriobasile.github.io/

outline


Hate Speech in Social Media
Hate Speech Detection
Data
Experiments
Machine Learning for NLP

hate speech in social media

"(language that is) abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination.

It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth"

(Erjavec and Kovacic, 2012)

hate speech in social media

Racism

Misogyny/sexism

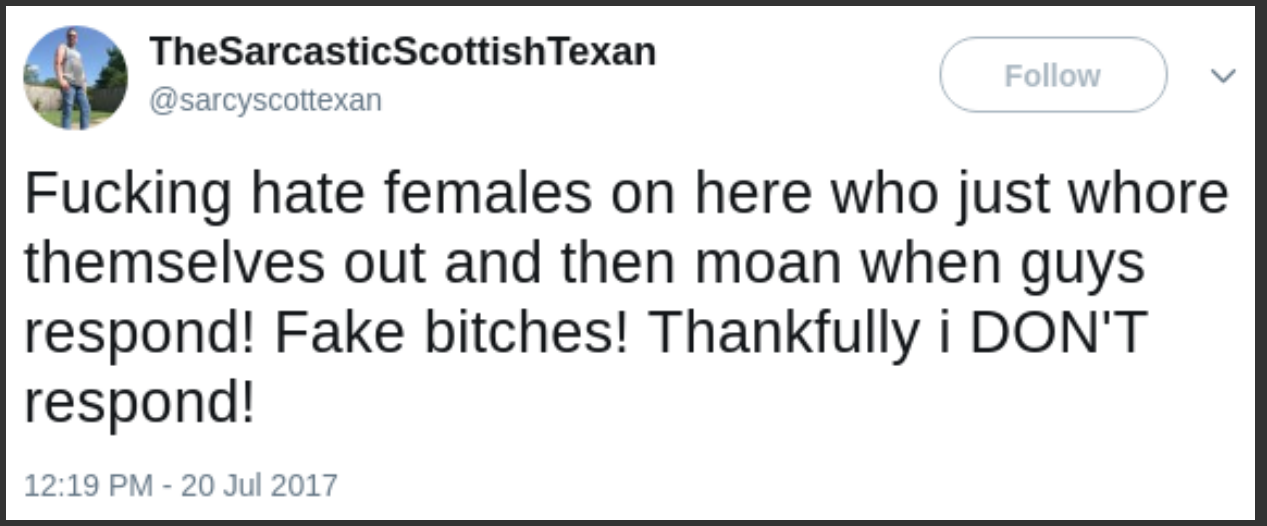
Homophoby

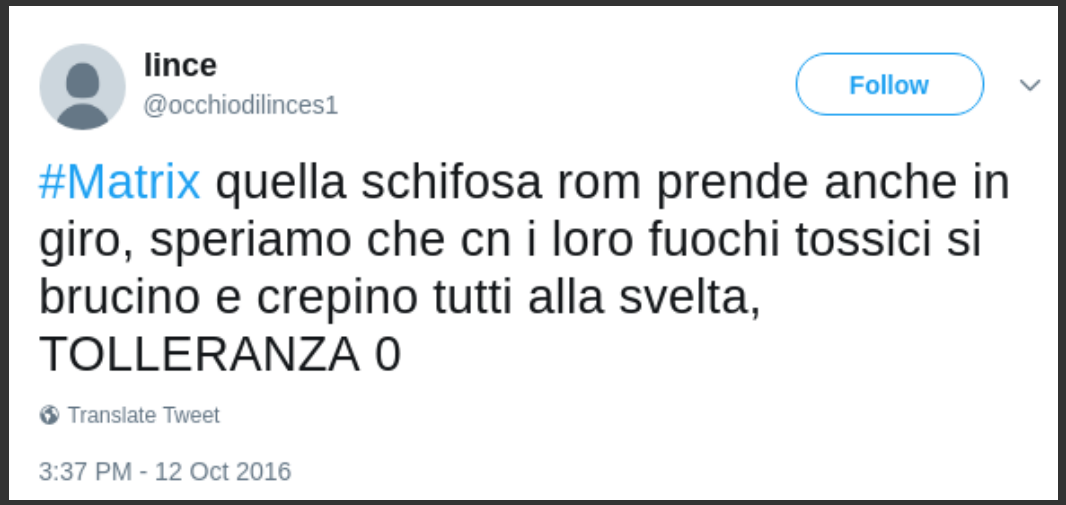
...

Related to:

Sentiment Analysis
Stance Detection
Irony and Sarcasm
Fake news, trolls, rumours, terrorism...

hate speech in social media





Demos (UK)
2014: 10,000-15,000 racist messages
2016: More than 200,000 sexist messages
DAILY

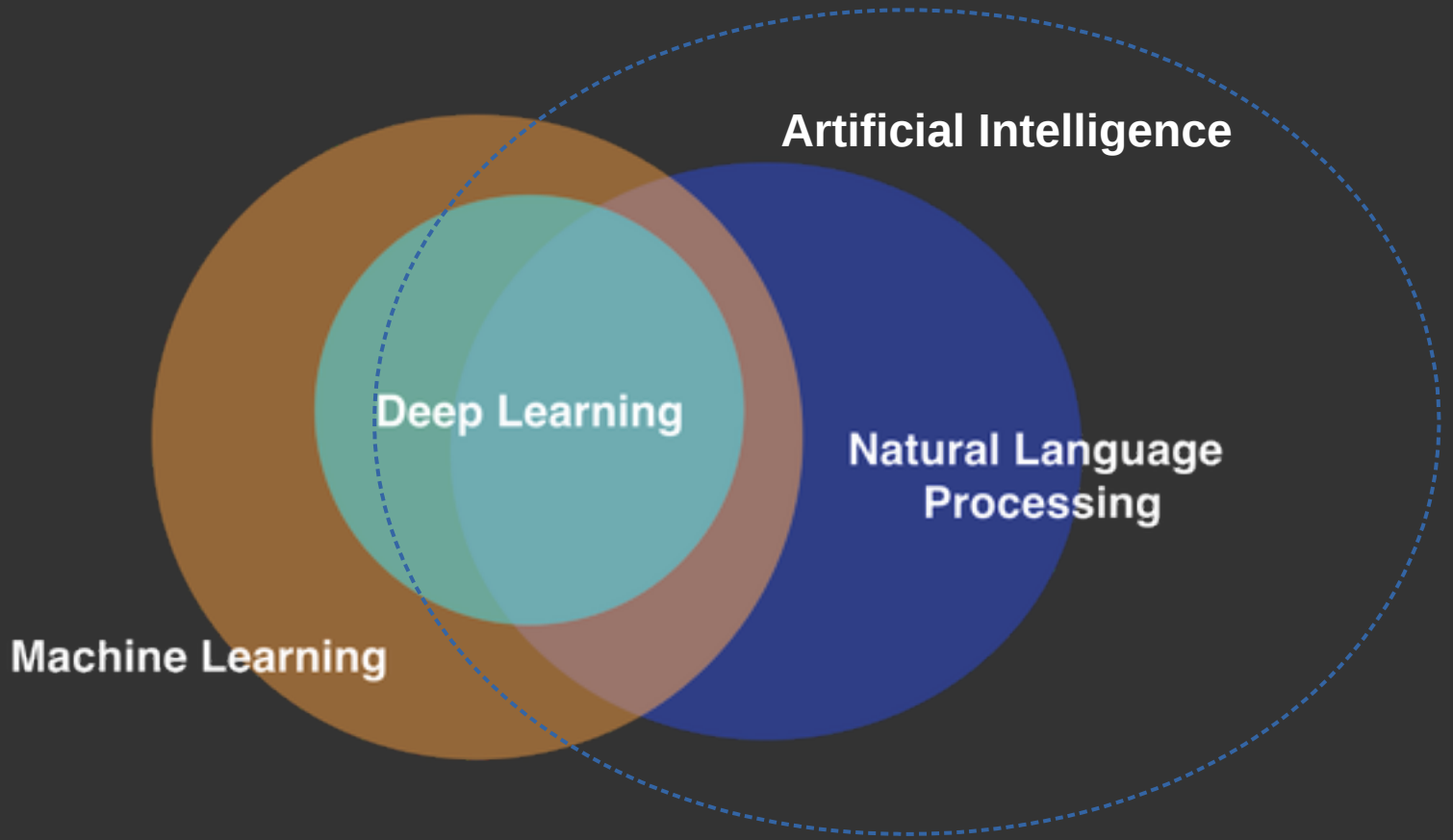
natural language processing

NOT (but ≃) Computational Linguistics

Morphology → how words are
Phonetics → how words sound
Syntax → how words relate
Semantics → what words means

Speech recognition
Machine Translation
Conversational agents
…

natural language processing



from quora.com

machine learning and NLP

Natural Language must be converted into numbers

Support from linguistic theories
(e.g. formal grammars, philosophy of language)

Supervised vs. unsupervised approach

TODAY: supervised text classification
training: text+labels → model
prediction: text+model → labeled text

machine learning and NLP

Example: Support Vector Machine with Bags of Words

"I really like Bologna" → POSITIVE (1)

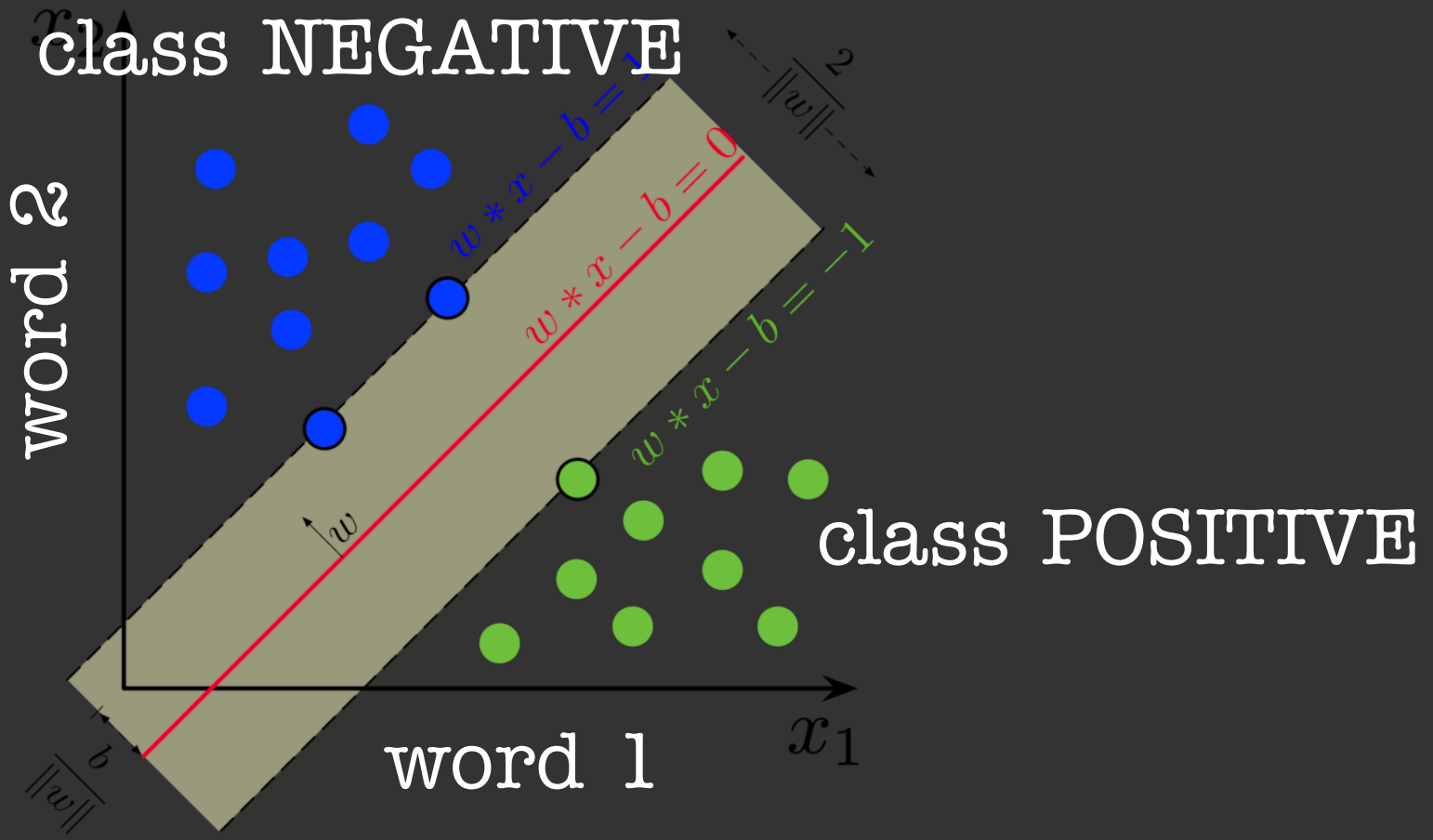
...	like	...	Bologna	...	class
...	1	0	1	...	1

...or with TF-IDF

...	like	...	Bologna	...	class
...	0.1	0	0.5	...	1

machine learning and NLP

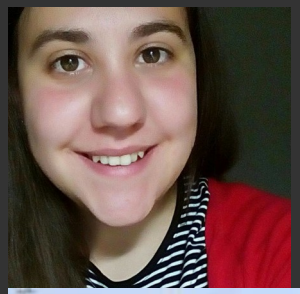
Example: Support Vector Machine with Bags of Words



machine learning and NLP

Example: Support Vector Machine with Bags of Words

14-ExLab@UniTo:
Automatic Misogyny Detection at IberEval 2018
1st place on English (91.3% accuracy)
and Spanish (81.5% accuracy)

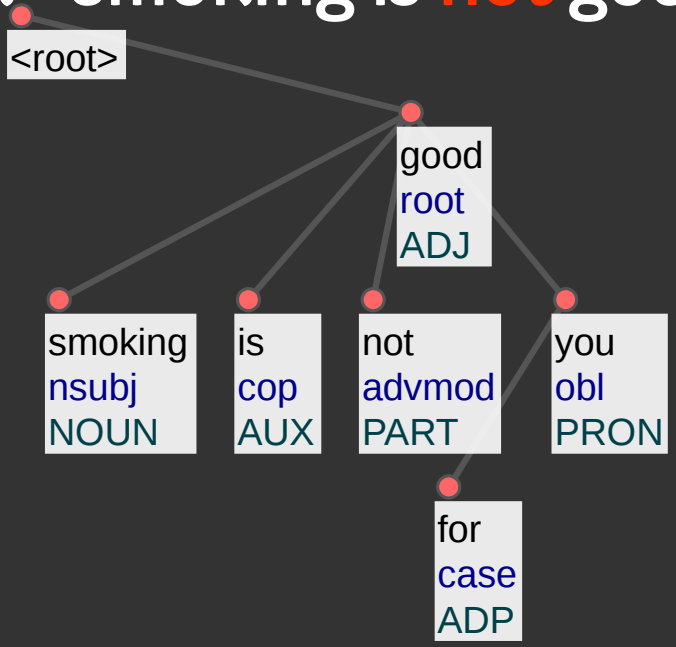


SVM with Bags of Words
+ Twitter-specific features
+ target-specific features

http://ceur-ws.org/Vol-2150/AMI_paper2.pdf

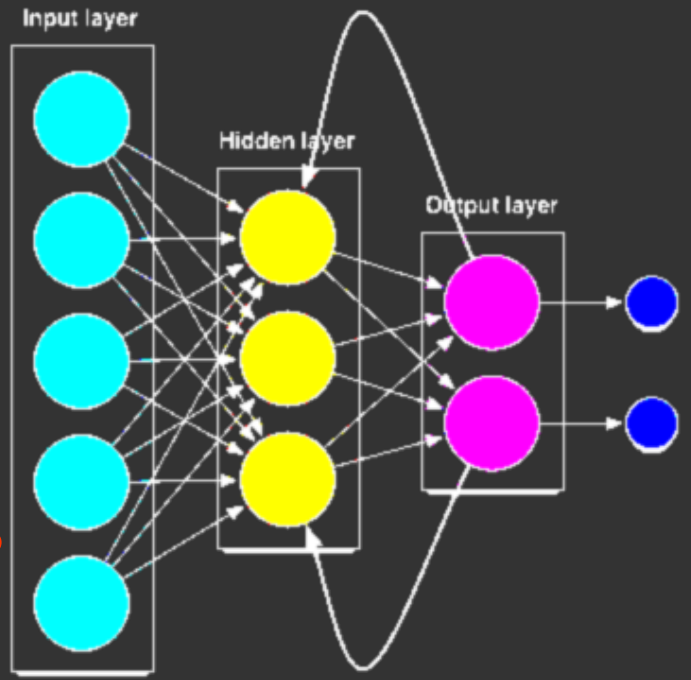
neural and deep

Words in natural language are not isolated.
e.g. "smoking is not good for you"



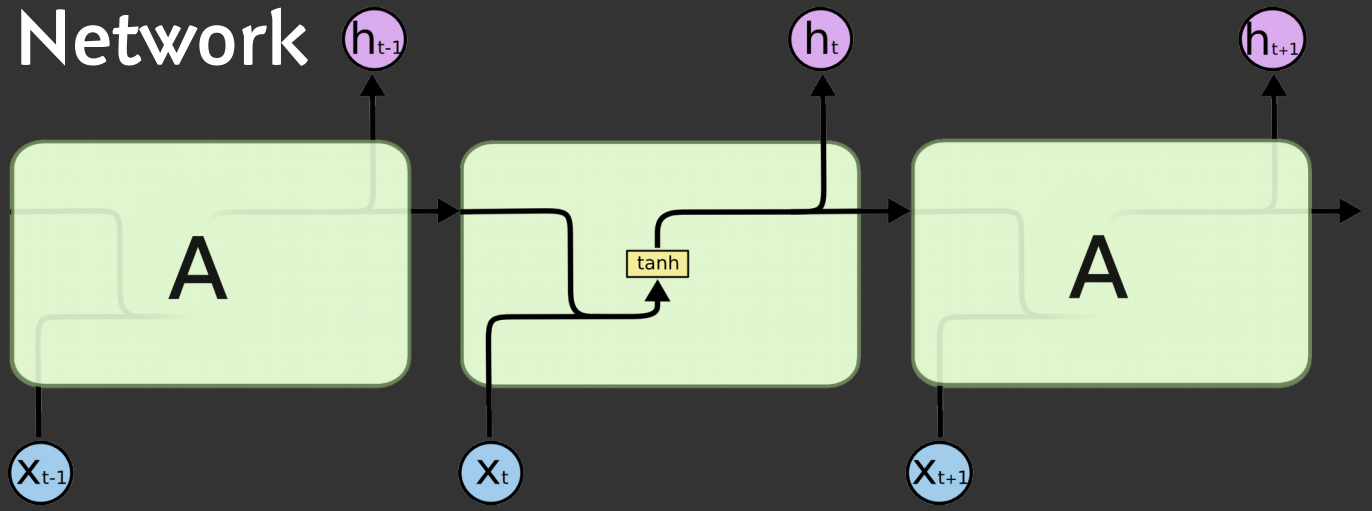
If words make features,
we need to model feature interaction

Natural language comes in sequences
→ recurrent architectures



neural and deep

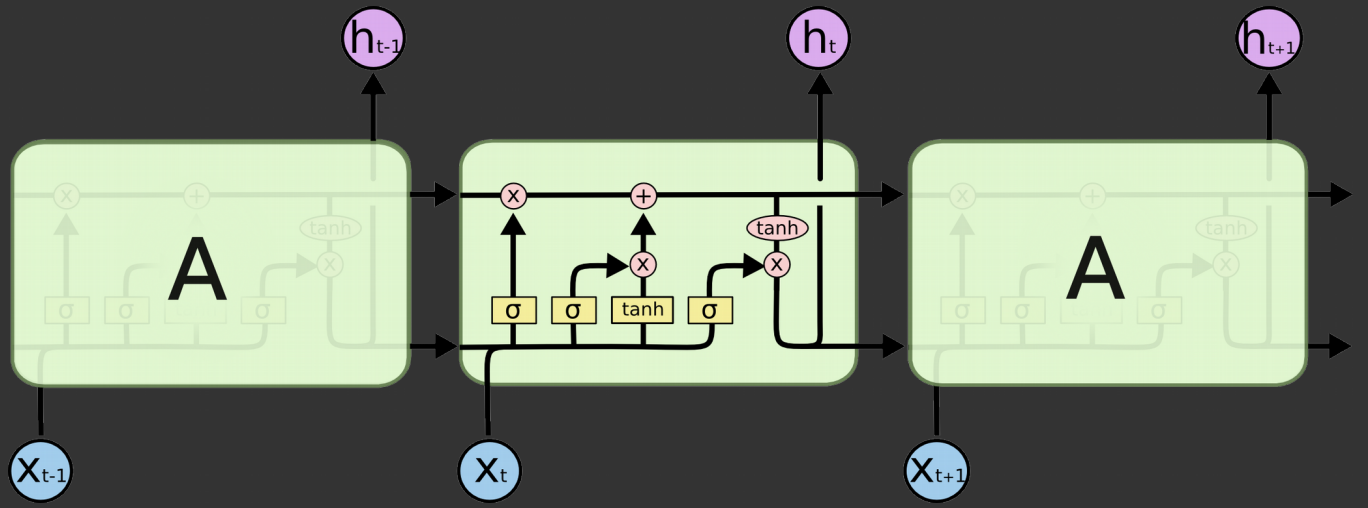
Recurrent Neural Network



Long Short-term Memory network

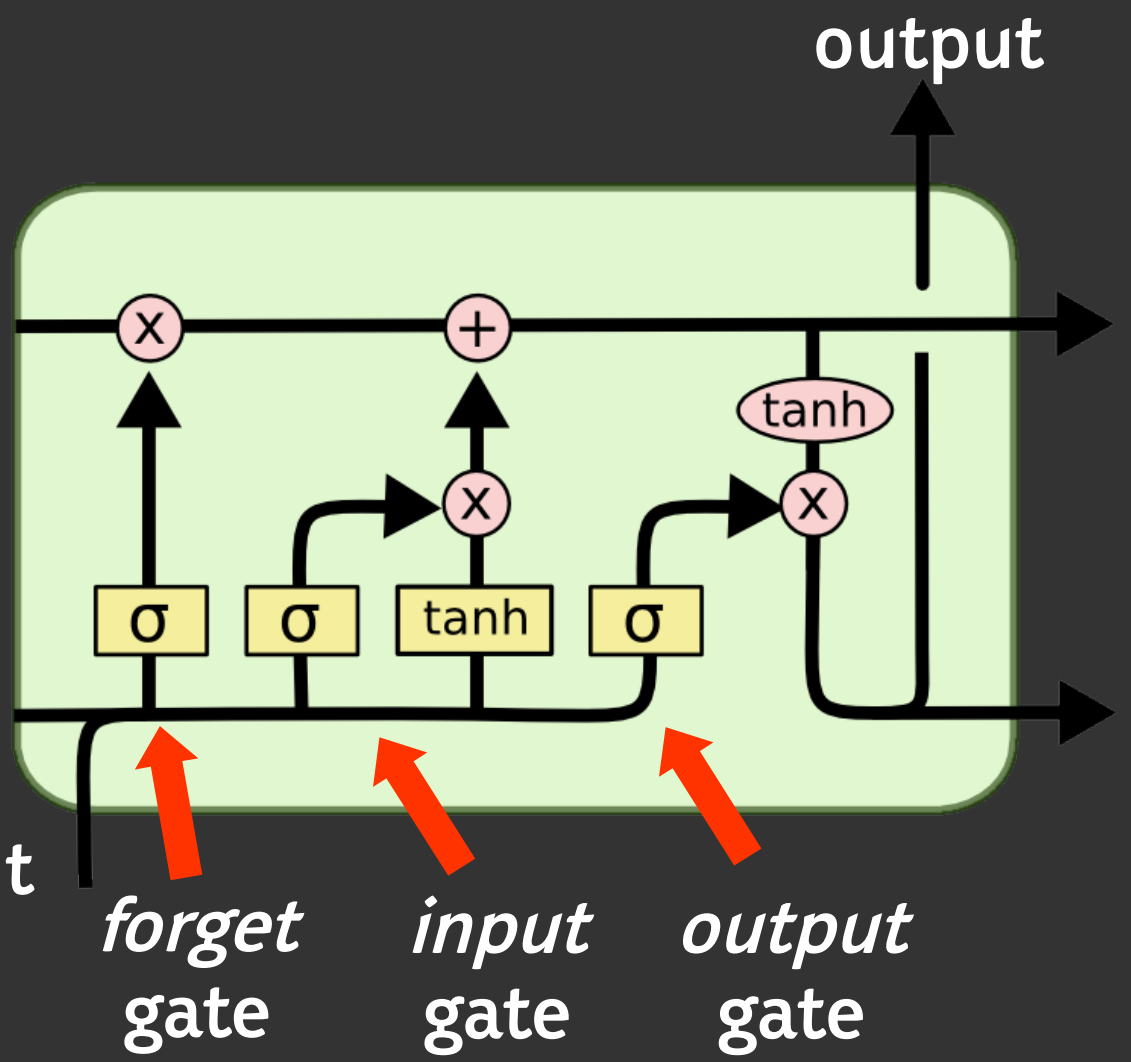
by Hochreiter & Schmidhuber (1997)

One word at a time!



neural and deep

LSTM unit

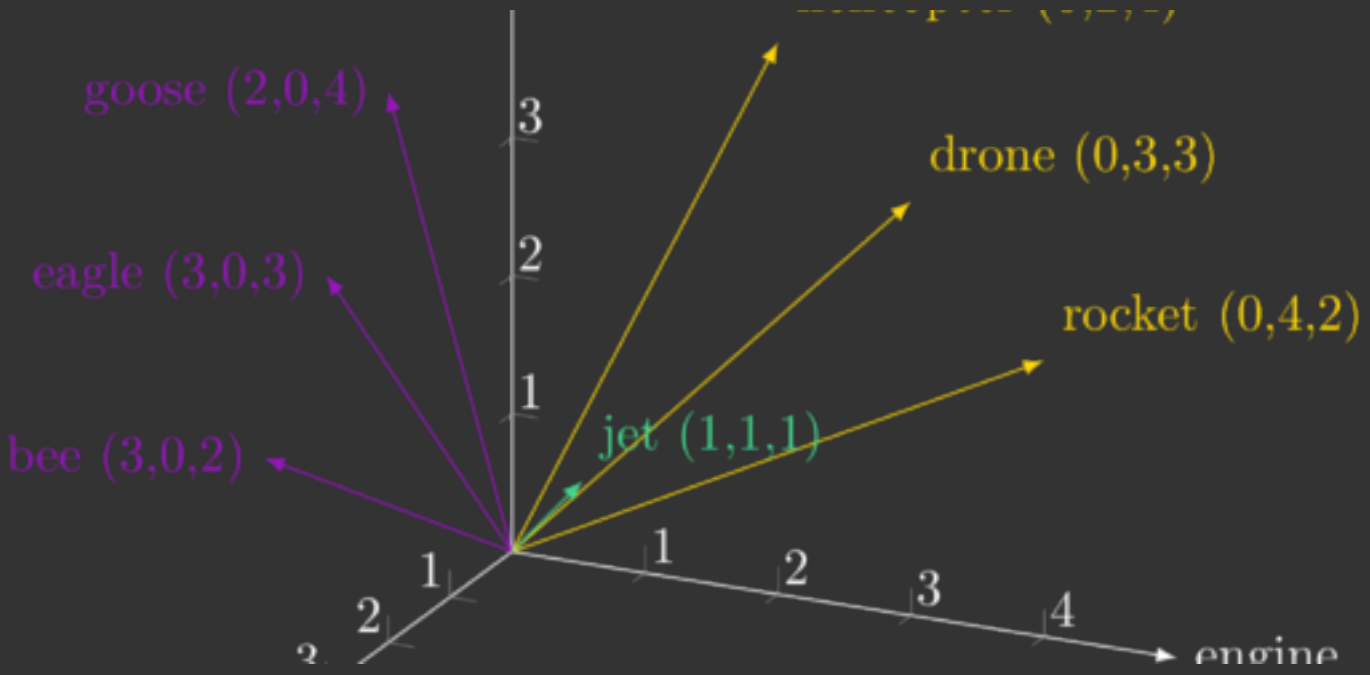


from Chris Olah's blog http://colah.github.io/

word embeddings

High-dimensional representations of words

Based on the distributional hypothesis
(Harris, 1954; Firth, 1957)



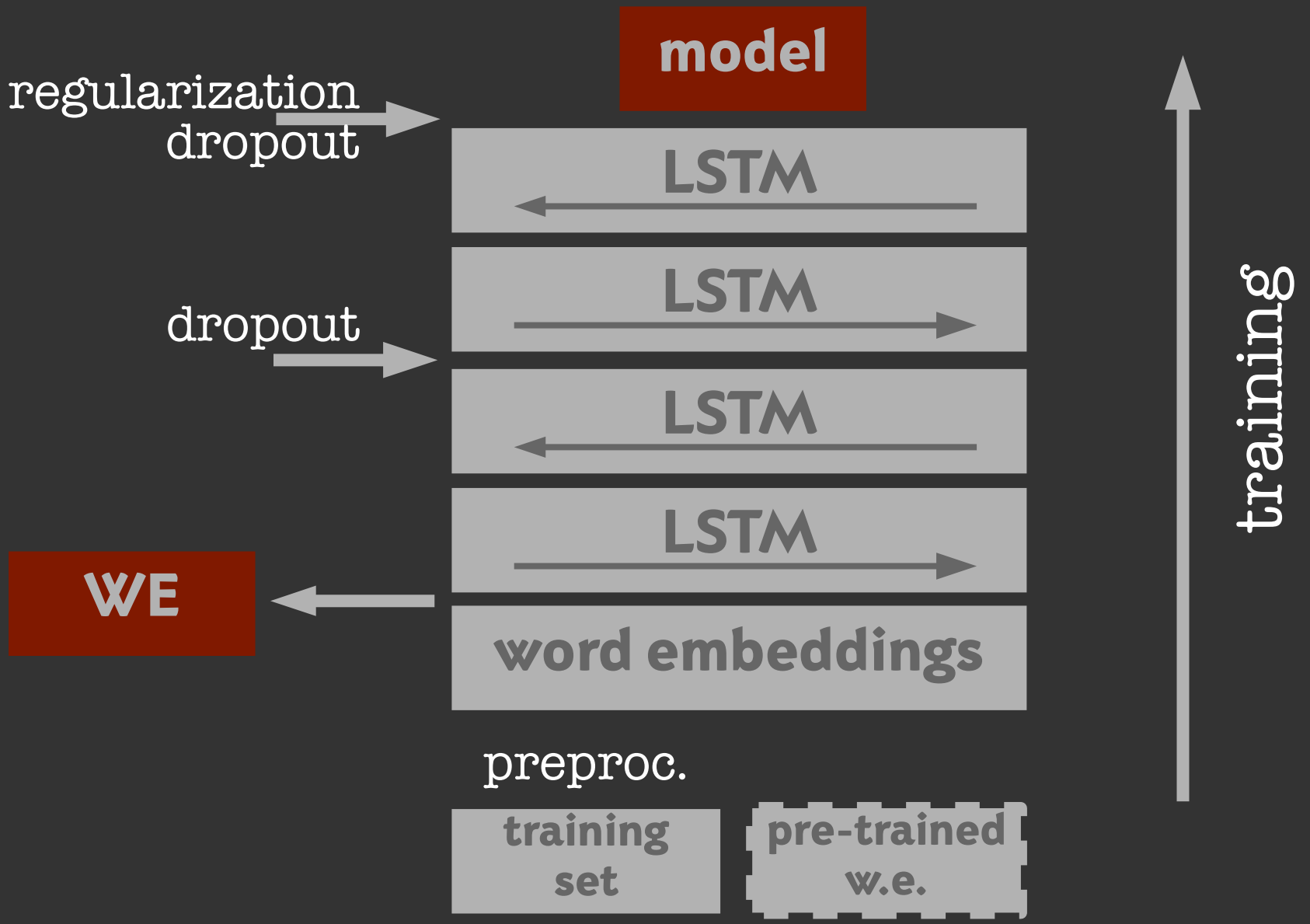
word embeddings

Pre-trained:
- Latent Semantic Analysis
- Random Indexing
- GloVe
- word2vec
- …

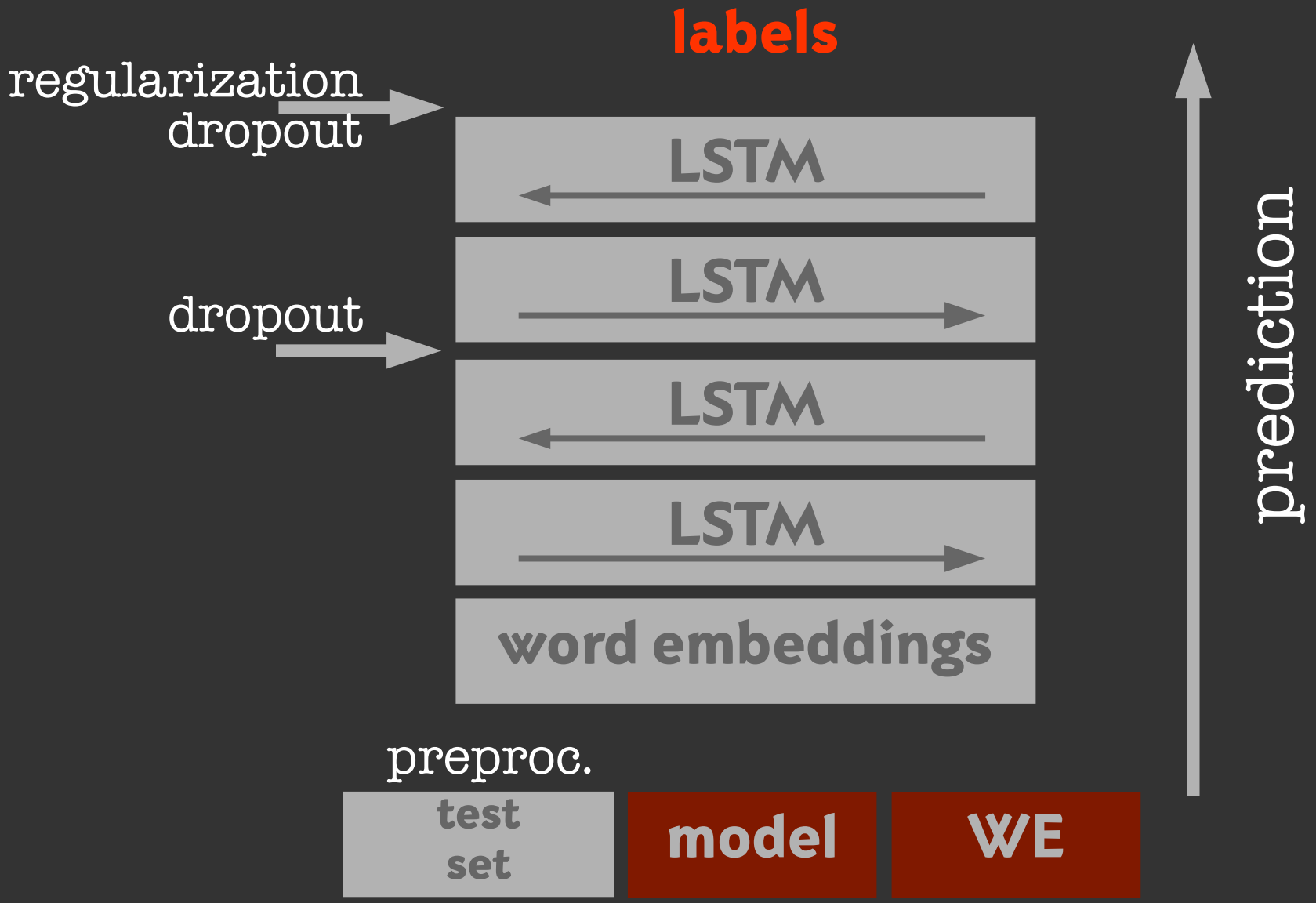
or let the NN learn them

```
modificò        0.23364 -0.32606 0.044304 -0.1217
umanesimo       0.48932 0.2108 0.2549 -0.15353 -0
pallavicini     0.14643 -0.6599 0.12375 -0.1158 0
primordiale     0.0048211 -0.099544 -0.20632 -0.1
legittimi       0.16595 0.057127 -0.026709 0.0361
controffensiva  0.29146 0.18019 0.01375 -0.25242
```

LSTM experiment



LSTM experiment



italian hate speech corpus

Corpus of hate speech on Twitter towards migrants and ethnic and religious minorities (Roma and Muslims in particular).

~6,000 tweets

Hate Speech Annotation: Analysis of an Italian Twitter Corpus. Clic-It 2017
http://ceur-ws.org/Vol-2006/paper024.pdf

An Italian Twitter Corpus of Hate Speech against Immigrants. LREC 2018
http://www.lrec-conf.org/proceedings/lrec2018/pdf/710.pdf

italian hate speech corpus

id_str	target	hate speech	aggressiveness	offensiveness	irony	stereotype	intensity
782117718791221248	ethnic group	no	no	no	no	no	0
782128837496745984	religion	no	no	no	no	no	0
782142959789670401	ethnic group	no	no	no	no	no	0
782145460664463360	Roma	no	no	no	no	0	
782165094318956548	ethnic group	no	weak	no	no	yes	0
782195284105371648	Roma	yes	no	strong	no	yes	1
782204731959734272	Roma	no	no	no	no	yes	0
782241280659169281	Roma	yes	strong	weak	no	yes	3
782268118194229248	Roma	no	no	no	no	no	0
782349137257922560	Roma	no	no	no	no	no	0
782462957842300930	ethnic group	no	no	no	no	no	0
782508027815485442	Roma	no	no	no	no	yes	0
782512181707440128	Roma	no	weak	no	no	no	0
782559406311477248	Roma	yes	weak	no	no	yes	2
782563896934666240	Roma	no	no	no	no	no	0
782584588103278597	ethnic group	no	strong	strong	no	yes	0
782588461006090240	religion	no	no	no	no	no	0
782596951283933184	religion	yes	weak	weak	no	yes	3
782614667759849472	ethnic group	yes	weak	no	yes	yes	3
782627058115641345	religion	yes	weak	no	no	yes	3
782640781290983424	ethnic group	no	no	no	no	no	0
782686657732640768	religion	yes	strong	no	no	yes	3
782787286857494528	ethnic group	no	no	no	no	no	0
782838281444683776	ethnic group	no	no	no	no	no	0
782838442044559361	ethnic group	yes	weak	weak	no	no	1
782861476126162944	religion	no	no	no	no	no	0

https://github.com/msang/hate-speech-corpus

Hurtlex

Multilingual lexicon of "words to hurt"

53 languages

17 categories + stereotype

asf	no	puss
cds	no	fiend
re	no	miscreants
ddp	yes	stupefy
an	no	diddlysquat
asf	no	boob
cds	no	fib
cds	no	puke
pr	no	streetwalker
re	no	terraist
cds	no	police-man
cds	no	gangs
cds	no	hypersexuals
asf	no	imbecility
ps	yes	stupidhead
asm	no	sap

http://hatespeech.di.unito.it/resources.html

evaluation

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn}$$

$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

results

Class 0 (no HS)

pre-trained emb.	precision	recall	F1 score
no	.911	**.933**	**.922**
FastText	**.925**	.891	.908
FastText+Hurtlex	.924	.914	.919

Class 1 (HS)

pre-trained emb.	precision	recall	F1 score
no	**.688**	.595	.630
FastText	.585	**.681**	.629
FastText+Hurtlex	.637	.665	**.651**

pre-trained emb.	macro-F1	accuracy
no	.776	**.871**
FastText	.769	.852
FastText+Hurtlex	**.785**	.868

a sentence classification library

Built on top of Keras and SpaCy

Multi-layer perceptron, deep NN, (Bi)LSTM

Support for pre-trained word embeddings

Regularization, dropout, early stopping, class weights

(NEW) Attention

https://github.com/valeriobasile/dnnsentenceclassification

a sentence classification library

conf.py

```
experiments = {
    "ihsccross": {
        "data": "ihsctrain",
        "language": "it",
        "model": "lstm",
        "preprocess": "lemma",
        "wordrepresentation": "embedding",
        "embedding_dimension": 300,
        "embedding_file": "wiki.it.vec",
        "max_length": 200},
    ...
}
```

https://github.com/valeriobasile/dnnsentenceclassification

evaluation campaigns

SemEval

Formerly SensEval

International evaluation campaign

ACL community

Since 1998

http://alt.qcri.org/semeval2019

evaluation campaigns

SemEval

Frame semantics and semantic parsing

- Task 1: Cross-lingual Semantic Parsing with UCCA [mailing list] [email organizers]
- Task 2: Unsupervised Lexical Semantic Frame Induction [mailing list] [email organizers]

Opinion, emotion and abusive language detection

- Task 3: EmoContext: Contextual Emotion Detection in Text [discussion group] [email organizers]
- Task 4: Hyperpartisan News Detection [mailing list] [email organizers]
- Task 5: HatEval: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter [mailing list] [email organizers]
- Task 6: OffensEval: Identifying and Categorizing Offensive Language in Social Media [mailing list] [email organizers]

Fact vs fiction

- Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours [discussion group] [email organizers]
- Task 8: Fact Checking in Community Question Answering Forums [mailing list] [email organizers]

Information extraction and question answering

- Task 9: Suggestion Mining from Online Reviews and Forums [mailing list] [email organizers]
- Task 10: Math Question Answering [mailing list] [email organizers]

NLP for scientific applications

- ~~Task 11: Normalization of Medical Concepts in Clinical Narrative [mailing list] [email organizers]~~
- Task 12: Toponym Resolution in Scientific Papers [mailing list] [email organizers]

evaluation campaigns

EVALITA

Evaluation campaign of language technology on Italian

AILC community

Since 2007

http://www.evalita.it/
http://www.ai-lc.it/

evaluation campaigns

EVALITA

Affect, Creativity and Style

- ABSITA - Aspect-based Sentiment Analysis (V. Basile, P. Basile, D. Croce, M. Polignano)
- ITAMoji - Italian Emoji Prediction (F. Barbieri, F. Ronzano, F. Chiusaroli, E.W. Pamungkas, V. Patti)
- IronITA - Irony Detection in Twitter (A. Cignarella, S. Frenda, C. Bosco, V. Patti, P. Rosso)
- GxG - Cross-Genre Gender Prediction (F. Dell'Orletta, M. Nissim)

Dialogue Systems

- iLISTEN - itaLIan Speech acT labEliNg (N. Novielli, P. Basile)
- IDIAL - Italian DIALogue systems evaluation (F. Cutugno, A. Origlia, M. di Maro, C. Tortora, B. Magnini, M. Guerini, S. Falcone)

Hate Speech

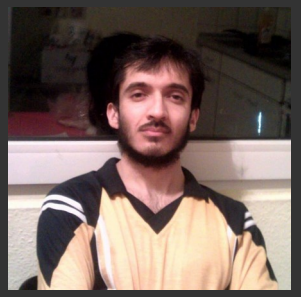
- AMI - Automatic Misogyny Identification (M. Anzovino, E. Fersini, P. Rosso)
- HaSpeeDe - Hate Speech Detection (C. Bosco, F. Dell'Orletta, M. Sanguinetti, F. Poletto, M. Stranisci, M. Tesconi)

Semantics4AI

- NLP4FUN - Solving language games (P. Basile. M. de Gemmis, G. Semeraro, L. Siciliani)
- SUGAR - Spoken Utterances Guiding chef's Assistant Robots (F. Cutugno, M. Di Maro, A. Origlia, C. Tortora)

MOAR!

Controversiality and opinion polarization



Demographic data analysis

Counternarratives against homophoby

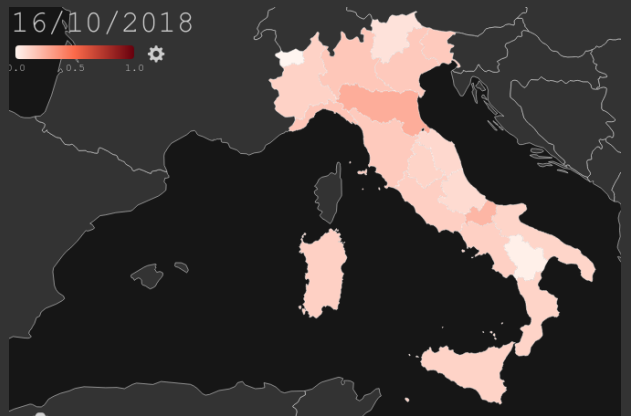
ARCIGAY
Associazione LGBTI Italiana

HATE SPEECH AND SOCIAL MEDIA

I HATE PREJUDICE

contro l'odio

http://mappa.controlodio.it/



ask me ask me ask me

About data
About annotations
About software
About projects
About (academic) life