monitoring vaccine confidence (with deep learning): the VCM platform

m. cristoforetti
joint work with L. Coviello and C. Furlanello
special thanks to Dr. A. Tozzi and A. D'Ambrosio (OPBG Rome)

DELVE2016
AMSTERDAM, SEPTEMBER 20TH 2016

FONDAZIONE BRUNO KESSLER

sponsored by CCM 2014 in collaboration with:
Ospedale Pediatrico Bambin Gesù - Roma, Regione Veneto,
Regioni Liguria, Toscana, Emilia Romagna, Lazio, Marche, Puglia, Sardegna e Sicilia

Bambino Gesù
OSPEDALE PEDIATRICO

VACCINATION CONFIDENCE

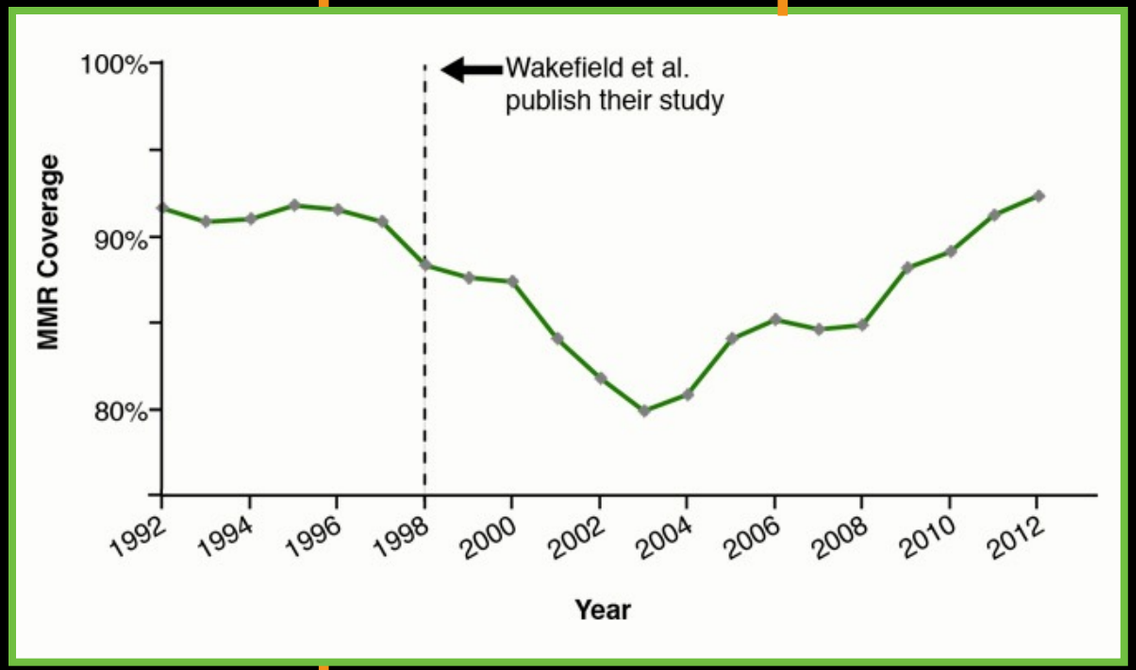
in the last years numerous events highlighted
a decrease in the **confidence** on **vaccination** programs

a crucial role in this phenomenon has been played by **mass media** and the **web spreading news** without any scientific foundation feeding media plagues

it is therefore relevant **monitoring confidence** of the public and their information needs on vaccination as emerging from the analysis of the **material** available **from the internet**

VACCINATION CONFIDENCE
a case study

The Lancet publishes a paper by Wakefield et al. titled "Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children." A press conference and interview in which Wakefield says it is his "*feeling that the… risk of this particular syndrome developing is related to the combined vaccine, the MMR, rather than the single vaccines*" set off a media storm.



The study involved just 12 children, would turn out to have some serious flaws and even to contain apparently falsified data.

1997 1998 2004 2010 2013

In 1997, the year before the paper was published, measles vaccination rates in the United Kingdom were over 91%.

They started to fall in 1998 and in 2003-2004 reached a nadir of just 80%, although rates were even lower than that in specific areas.

Of the 13 authors on the 1998 Wakefield paper, ten formally retract its interpretation, stating a wish to make clear that the paper established "no causal link" between the MMR vaccine and autism

The Lancet, after an extensive investigation, including investigative work by journalist Brian Deer, issues a formal retraction of the 1998 Wakefield paper

Only in recent years have MMR vaccination rates started climbing again in the U.K., reaching about 90% in 2013.

VACCINATION CONFIDENCE MONITORING PLATFORM

a platform that scrapes the web in order to **monitor**, **collect** and **analyse** articles, web pages, social posts and videos regarding specific topics

topic can be whatever

vaccine and vaccination

VACCINATION CONFIDENCE MONITORING PLATFORM

vaccine and vaccination

monitor

online news

web page production

posted video

social media

collect

filter by topic / keywords / channels

analyse

topic discovery

sentiment analysis

influencer

new web sites

VCMP: monitor and collect

Initial setup by domain expert

Topic, keyword and channel selection

Train dataset

samples of pertinent and non pertinent search queries

labeled by experts

Scraper Modules

News/Blogs

bing

Search Results

YAHOO!

bing

Social

Videos

vimeo

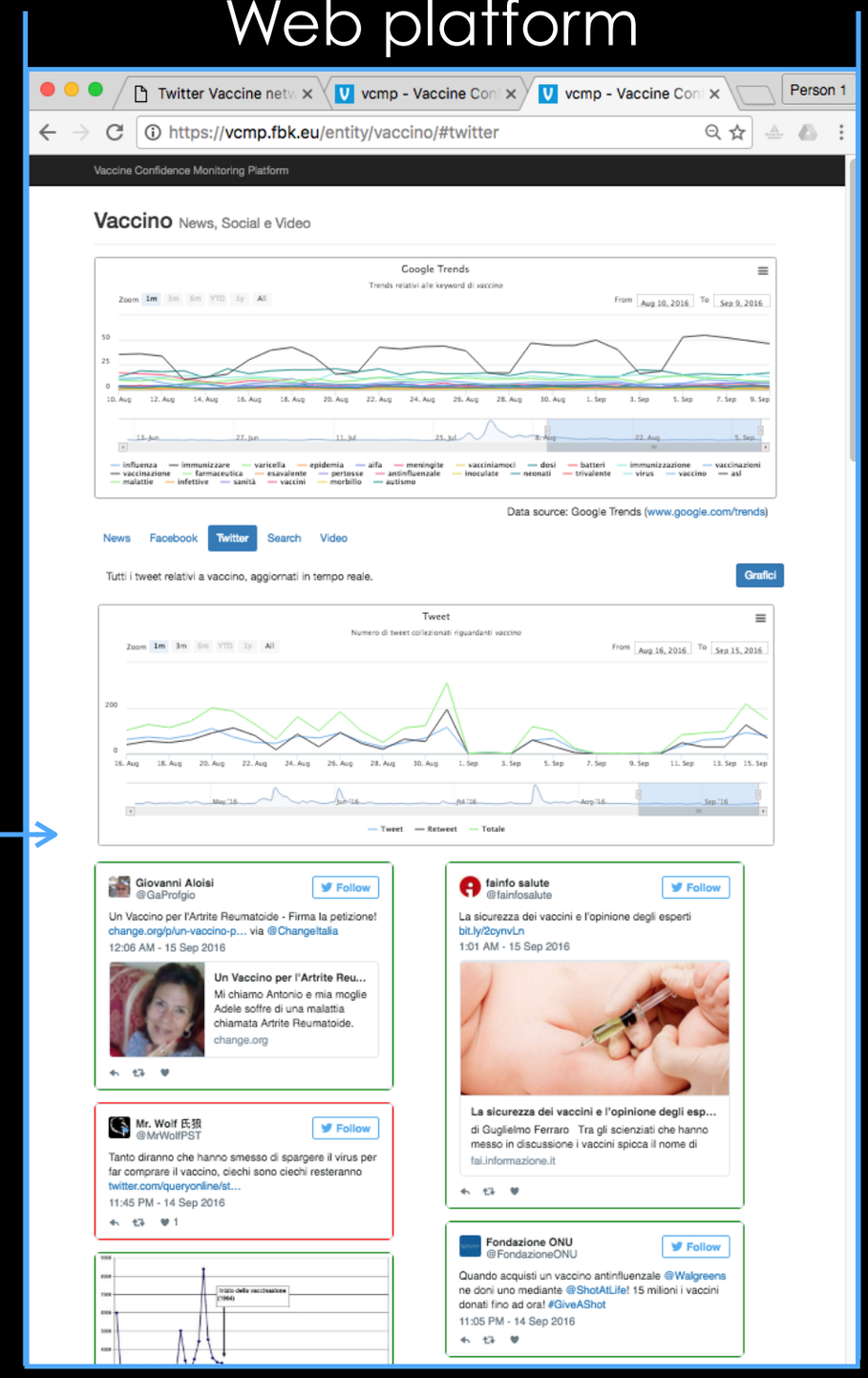
Pertinence filter

word2vec dictionary + Convolutional Neural network

Polite scraping

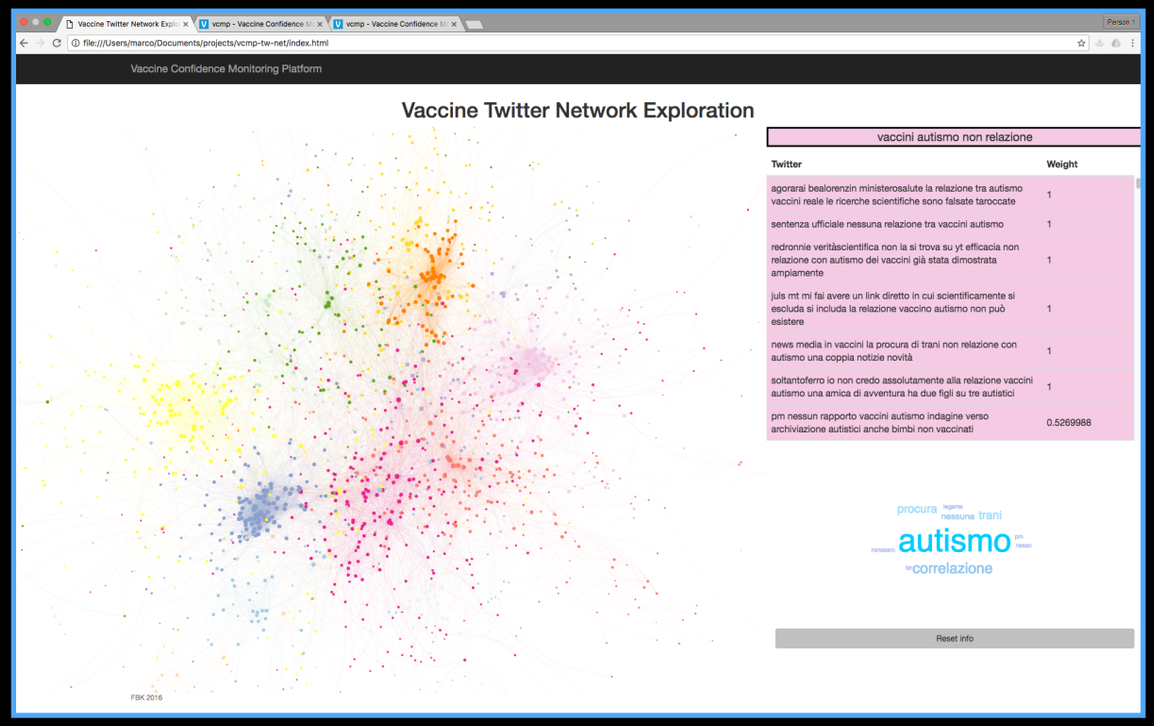
phantomJS + Selenium

Web platform

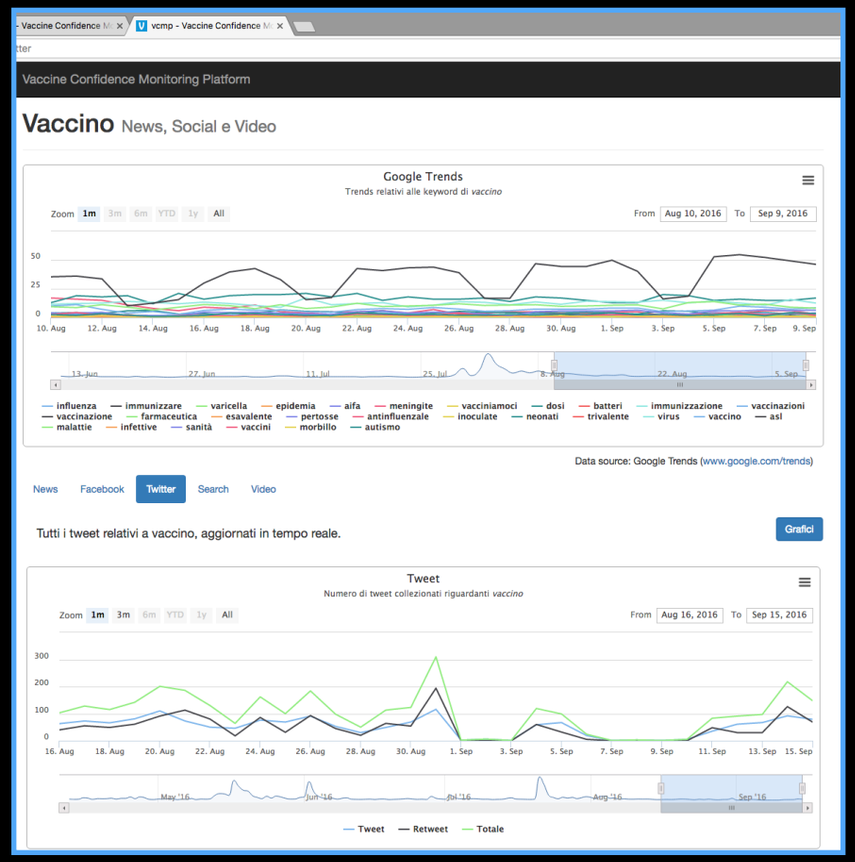


https://vcmp.fbk.eu

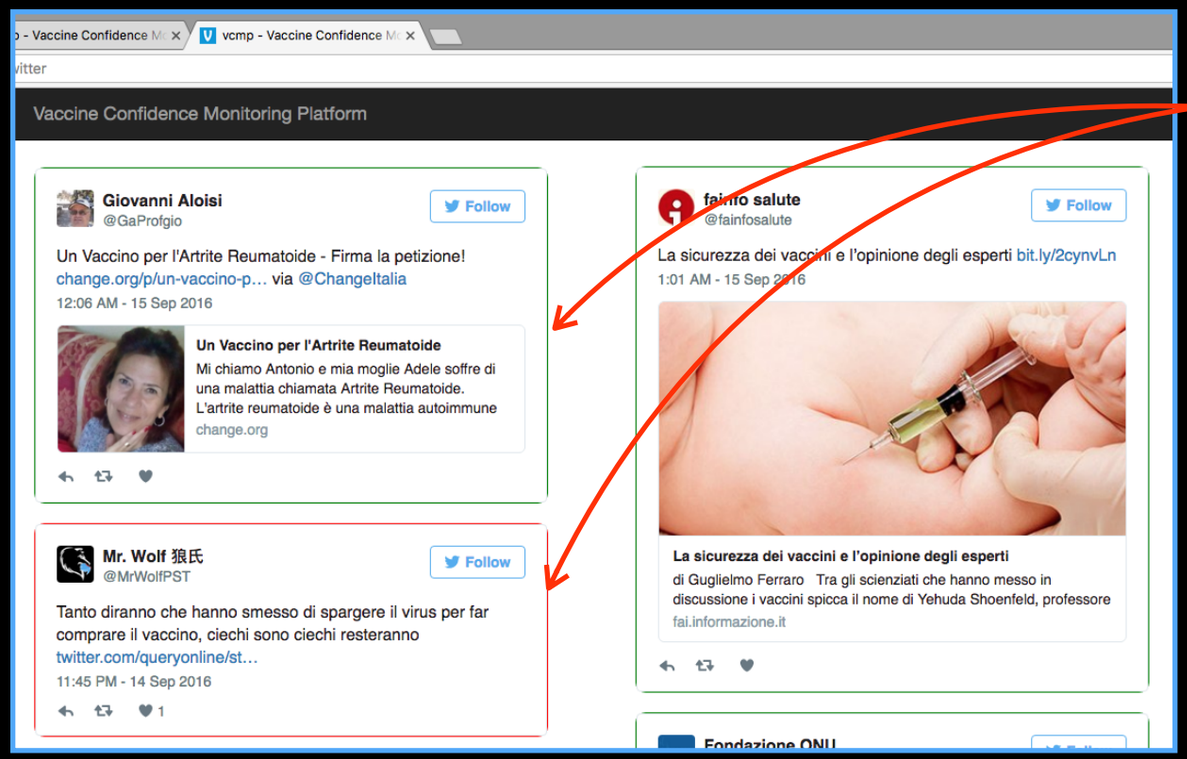
VCMP: analytics



Network analysis



Trends analytics

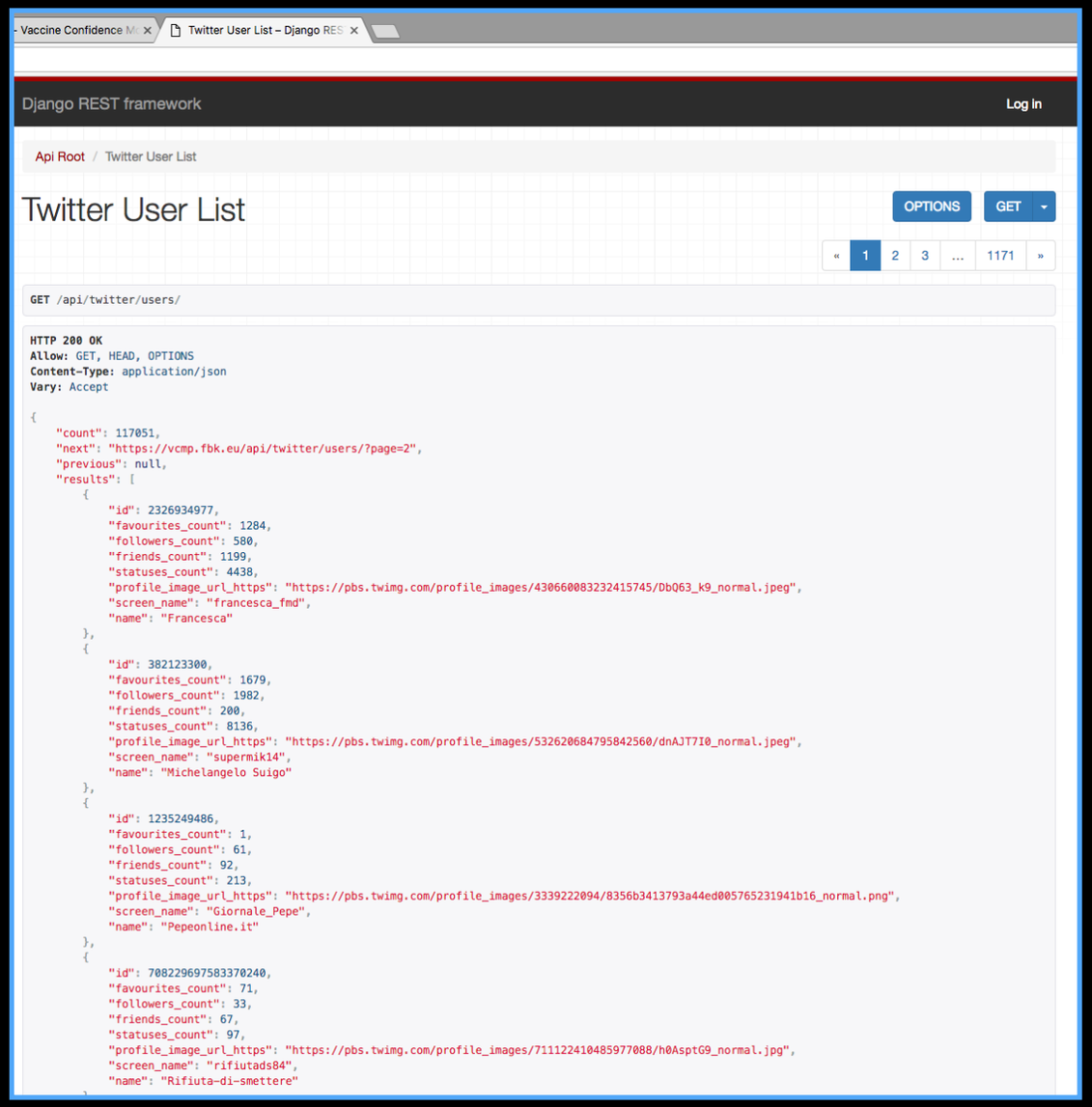


Sentiment analysis

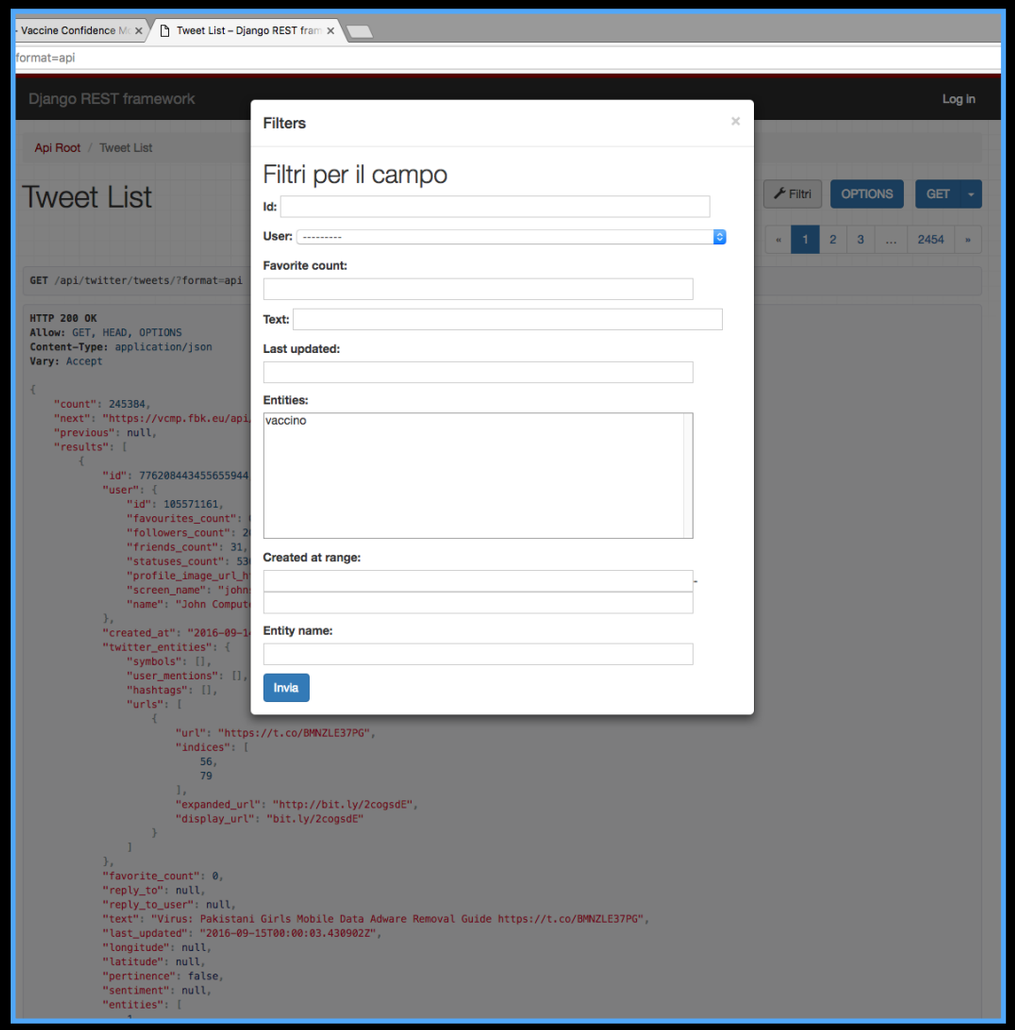
classification axis
and train dataset

decided and labeled
by experts

VCMP: extraction of data [api]

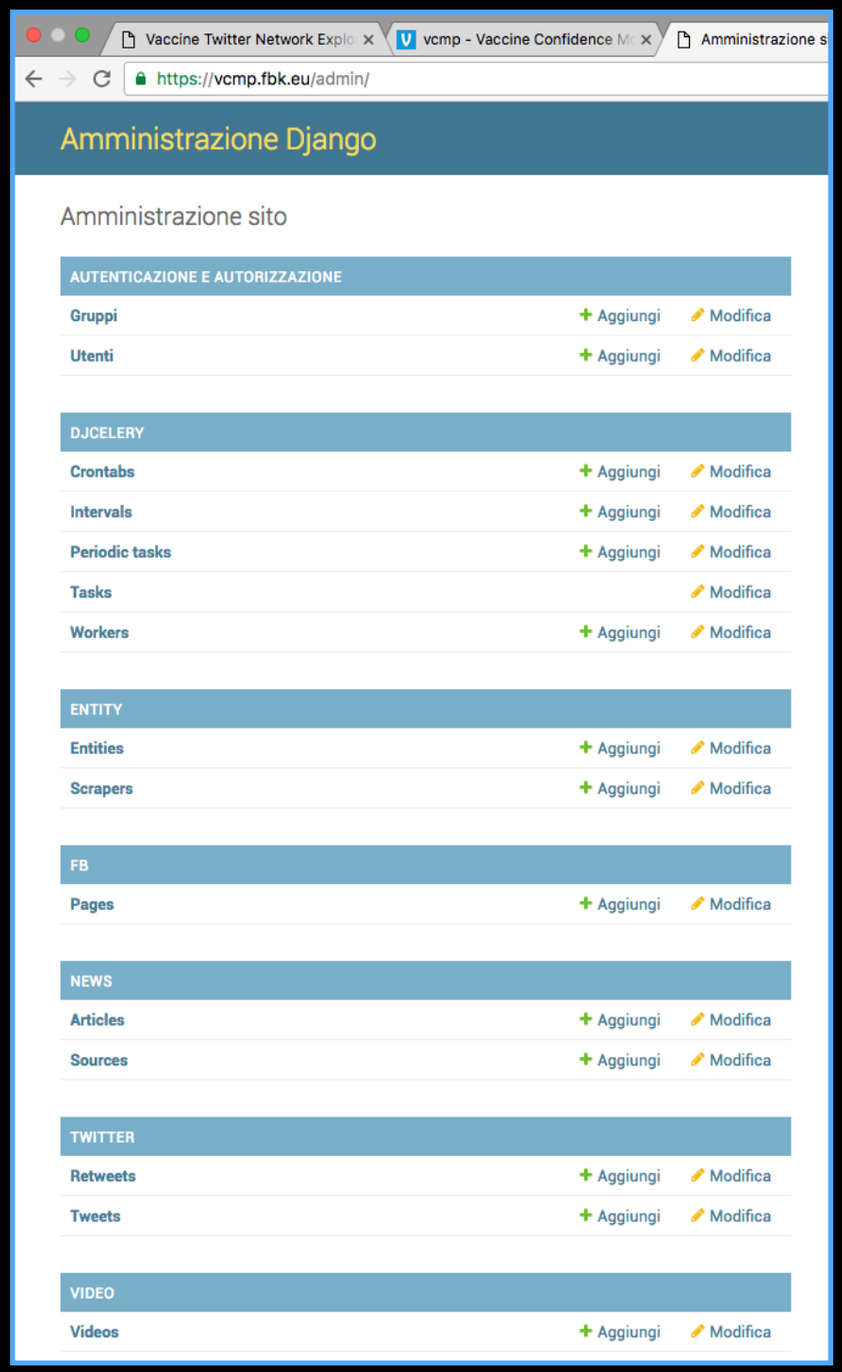


GET method to extract specific subsets of data



possibility to filter the results

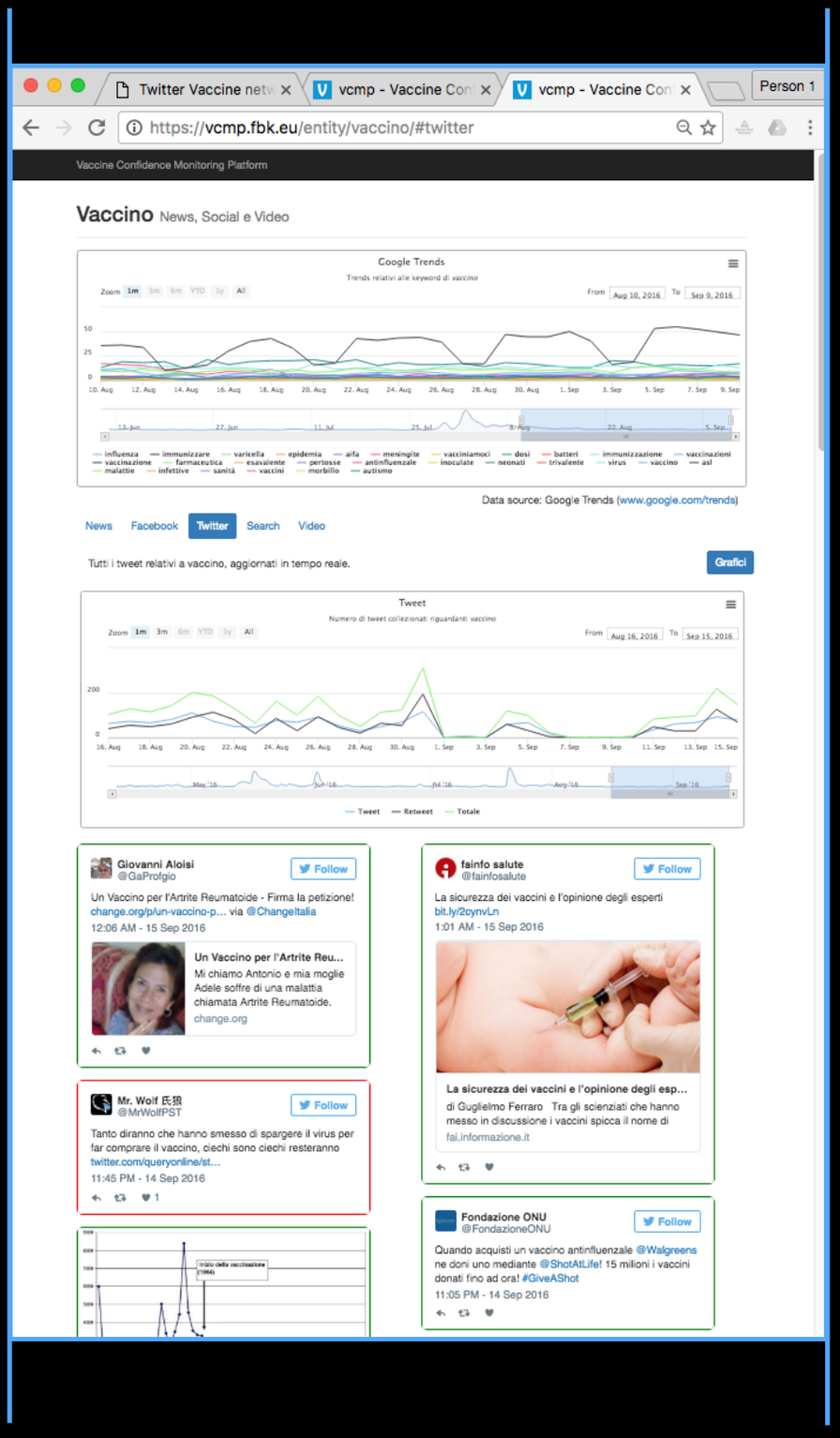
VCMP: administration and tuning



both in the configuration and production phase the platform algorithms can be easily tuned by expert of the field/topic

in this case [vaccine] by Medical Doctors

VCMP: infrastructure



Web platform

https://vcmp.fbk.eu

Django webserver
with PostgreSQL
and Celery

Backend

VCMP: dataset as at 14 September 2016

Collected documents since April 11, 2016

Source	Avg per day	Total
News	229	36125
Facebook	13	2119
Tweets	1450	229149
Retweets	1249	188562
Videos	54	8499

Collected articles for selected newspapers

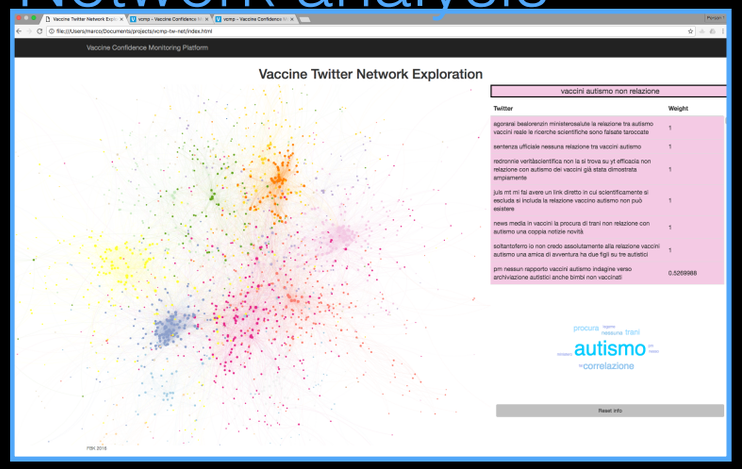
Source	Count
ANSA	672
Corsera	621
Repubblica	604
Il Sole 24 Ore	406
La Stampa	351
Il Giornale	183
Il Fatto Quotidiano	149

ANALYSIS IMPACT OF TV events
EXAMPLE: tweets during the TV broadcast “Presa Diretta” (10 January, 2016):
E. Agricola, A. D’Ambrosio, A. E. Tozzi, C. Rizzo, F. Gesualdo, E. Pandolfi, A. Ferro, A. Siddu, M. C., Luca. Coviello, C. Furlanello)
Monitoring Twitter streams to assess the impact of information on vaccines provided by a TV broadcast (SUBMITTED)

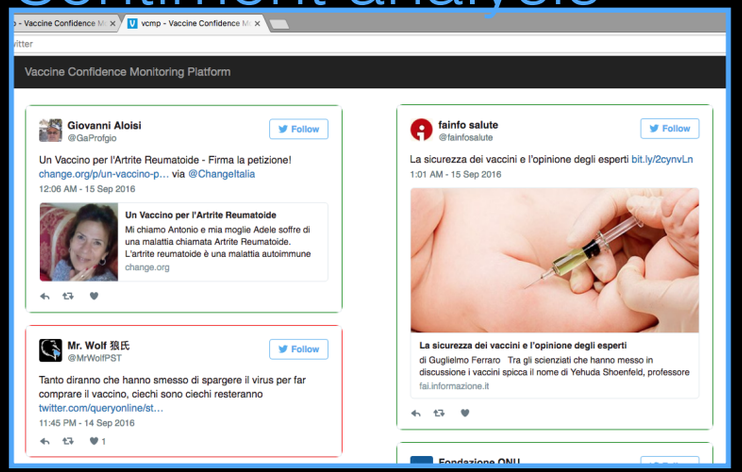
VCMP demo 1

VACCINATION CONFIDENCE with Deep Learning

Network analysis



Sentiment analysis



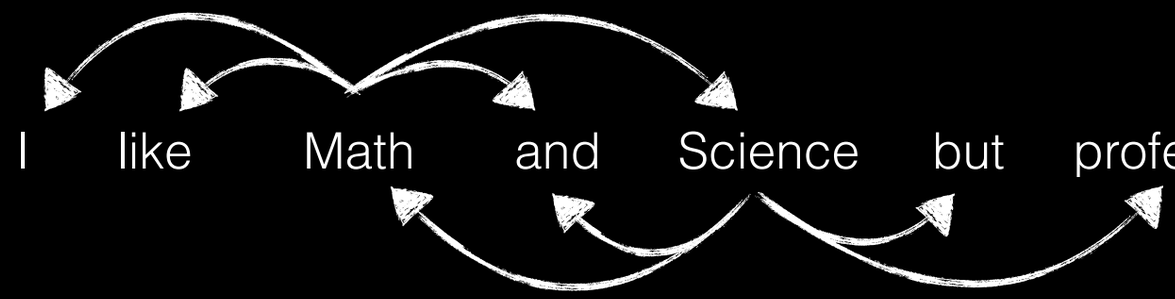
pertinence and sentiment analysis, exploration of the data collected and information extraction from the text, require application of NLP techniques

in the VCMP we apply state of the art
Deep Neural Network
to address this tasks

word2vec language model to capture similarities between words
Convolutional Neural Networks (CNN) for classification problems

word2vec: word embeddings

word2vec is a neural networks based model that learns word vector representations in an unsupervised manner
It tries to predict the surrounding words for every word



word2vec tries to maximize the log probability of any context (outer) word within a window given the current center (inner) word.

$$p(w_O|w_I) = \frac{\exp\left(v_{w_O}^t v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left(v_w^t v_w\right)}$$

Efficient estimation of word representations in vector space, Mikolov 2013

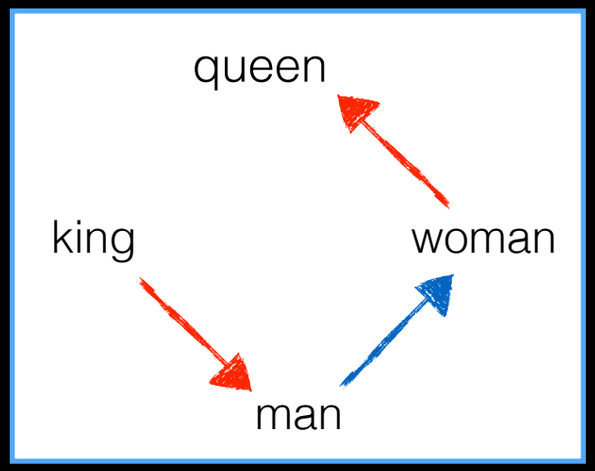
word2vec: word embeddings

word2vec returns low dimensional representations of words, which capture syntactic and semantic similarities
Similarities are represented through vector offsets

$X_{car} - X_{cars} \approx X_{family} - X_{families}$

$X_{shirt} - X_{clothing} \approx X_{chair} - X_{furniture}$

$X_{king} - X_{man} + X_{woman} = X_{queen}$



word2vec for Italian

The model has been trained on an Italian corpus composed of a Wikipedia dump, books, online newspaper articles with their articles and contents from selected websites extracted using commoncrowl.org
Compared with *Word Embeddings Go to Italy: a Comparison of Models and Training Datasets*, Berardi 2015

	Berardi et al.	VCMP
capital-common-countries	86.17%	89.5%
capital-world	55.61%	72.1%
currency	4.45%	4.7%
city-in-state	26.33%	37.1%
regione-capoluogo	32.16%	44.2%
family	63.00%	64.8%
adjective-to-adverb	15.16%	16.0%
opposite	17.48%	23.4%
comparative	4.17%	0.0%
superlative (absolute)	28.38%	29.2%
present-participle (gerund)	66.43%	73.6%
nationality-adjective	77.33%	86.4%
past-tense	40.48%	43.1%
plural	47.50%	51.0%
plural-verbs (3rd person)	78.97%	85.0%
plural-verbs (1st person)	22.79%	27.5%
present-remote-past-verbs (1st person)	17.41%	5.6%
masculine-feminine-singular	46.64%	50.6%
masculine-feminine-plural	16.55%	17.3%
overall accuracy	43.63%	50.7%

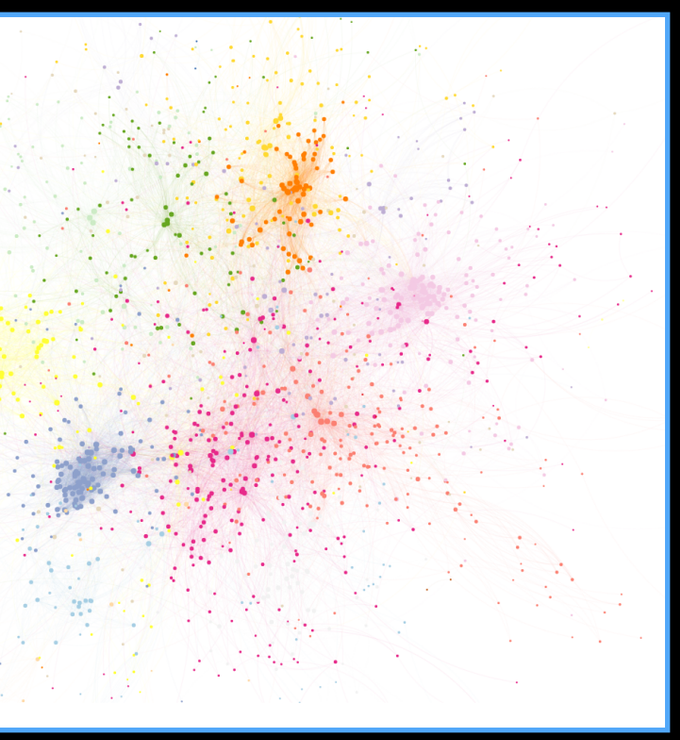
word2vec:
twitter network exploration

- Dictionary generated using word2vec with 20M tweets including the one collected from the VCMP
- Distance between tweets based on cosine-similarity in terms of the word2vec vectors

- Clustering using standard community detection algorithm

- Improvement with respect to a bag of words approach because tweets that use synonyms in terms of word2vec dictionary are close to each other

- Detection of emerging arguments. Temporal dynamics of clusters



most similar words to "meningite":

```
('vaccino', 0.6183883547782898),
('papillomavirus', 0.5869245529174805),
('profilassi', 0.5734522342681885),
('parotite', 0.5724117755889893),
('vaccinazione', 0.5576623678207397),
('vaccinazioni', 0.547511100769043),
('meningococco', 0.5446576476097107),
('poliomielite', 0.5434256196022034),
('vaccinarsi', 0.5358799695968628),
('antinfluenza', 0.5352928638458252)
```

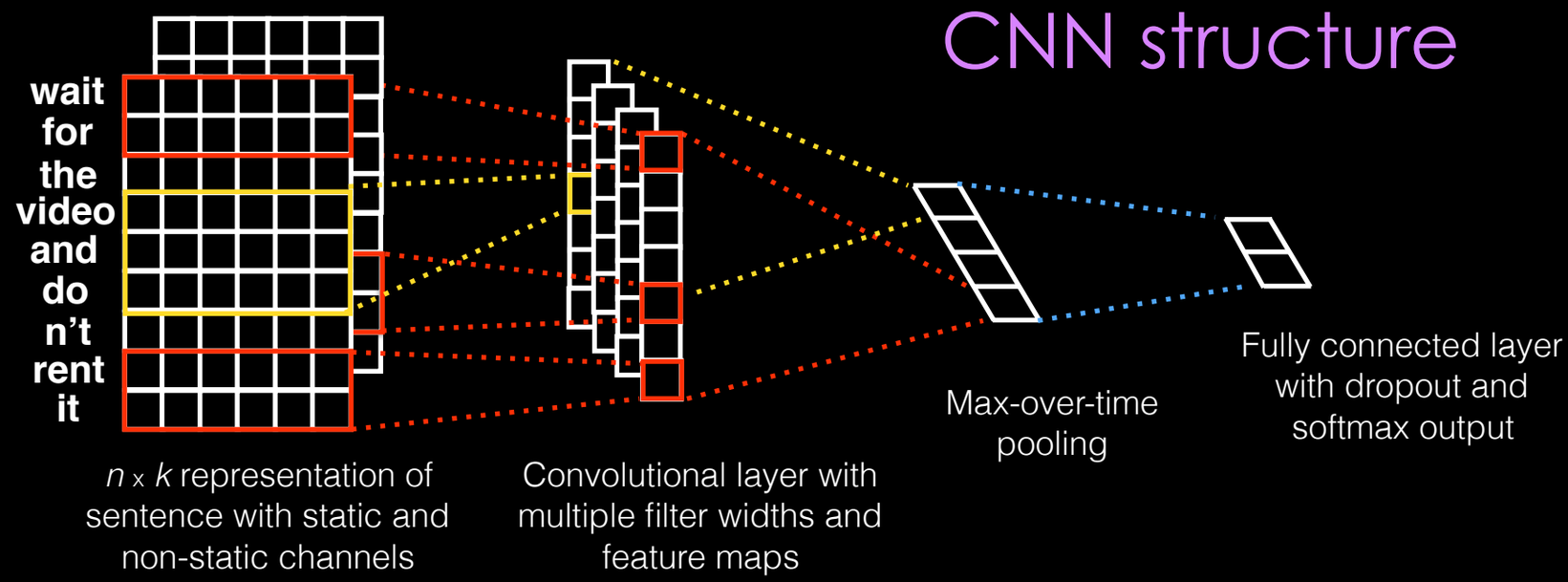
"vaccino" is excluded calculating the distance

pontedera contro la meningite scatta ancora il ~~vaccino~~ day

Twitter	Weight
ancora meningite io non mi sono ancora vaccinata	0.2575833
grande successo per vaccinoday pontedera in tanti per la vaccinazione contro la meningite	0.2479431
vaccino day contro la meningite ecco dove quando farlo vaccino salute	0.226333
~~vaccino~~ day immunizzati number elbani per il meningococco lunedì aprile number tirreno elba news	0.1928407
meningite incubo torna in toscana vaccino meningitec toscana careggi	0.02151801
prato news toscana post meningite nessuno si vaccina più asl alle famiglie prima di partire per le vaca	0.02111891
retweeted siciliatoday todaysicilia catania allarme meningite partono le vaccinazioni ma nessuno dice	0.001847088
meningite toscana perche sopra number anni la gente deve pagare il vaccino perche non gratis per tutti paghi ti arrangi	0.0006007928

CNN for sentence classification (Kim 2014)

Word vector representations are initialized with word2vec,
and they can be static or non static.
If the architecture is non static, the error is back-propagated into word vectors.
This helps learning task-specific word representations and similarity



Instead of having a fully connected neural network, CNNs apply filters
to every subregion of the input to extract new features
CNNs learn representation for every single subphrase, learning that
some combination of words are more informative than others

If the input sentence is *"the country of my birth"*, the CNN considers
*"the country, country of, of my, my birth, the country of, country of my,
of my birth, the country of my, country of my birth"*

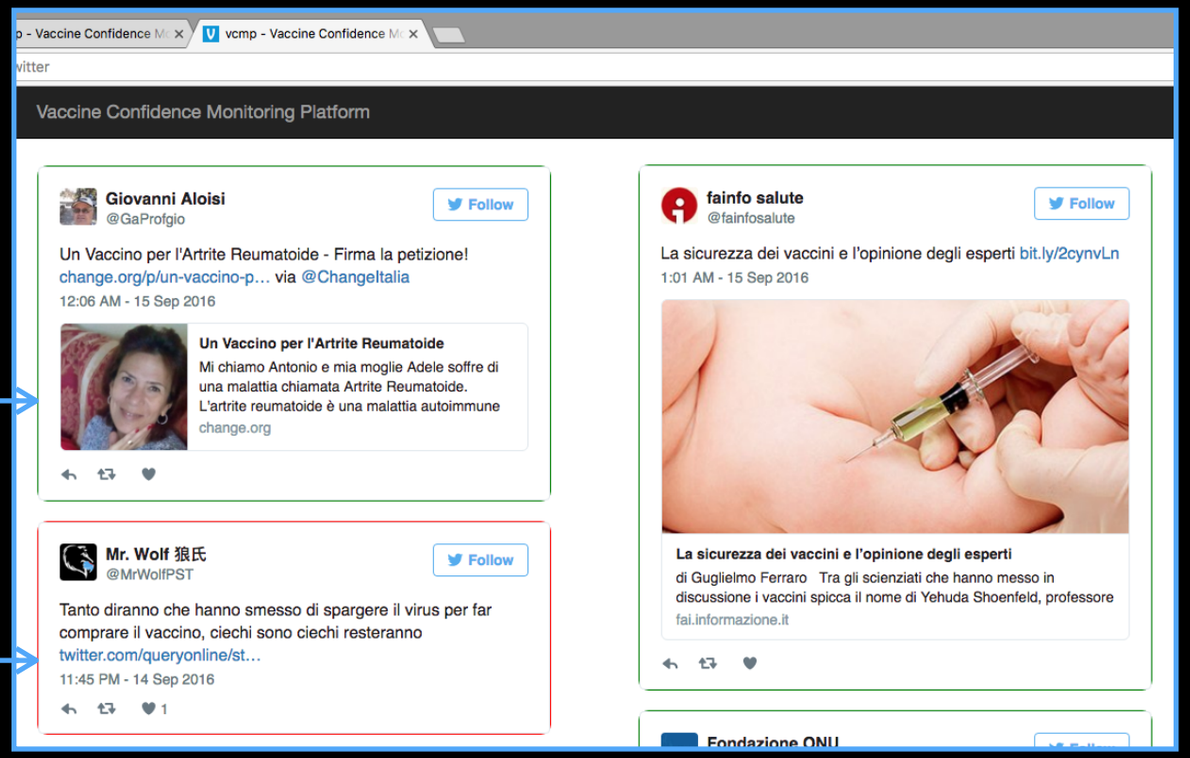
CNN for sentence classification
pertinence and confidence

First test with Tweets
2116 Tweets labelled by experts
In this first run the label was confident vs all the rest
(either against vax or non pertinent):
1528 positive tweets
588 negative tweets

word embeddings computed with word2vec
filter sizes: 3,4,5

5-fold cross validation: 78% accuracy

green border = pro
red border = against



VCMP demo 2

Acknowledgement

the CCM and OPBG

Dott. A. Tozzi

Dott. A. D'Ambrosio

the MPBA group in FBK

C. Furlanello

L. Coviello

A. Gobbi