Cos'è un lessico affettivo
L'esperienza di Sentix

Valerio Basile
Università degli Studi di Torino

CREA, Roma
19/2/2020

Ricercatore @ Università degli Studi di Torino
Dipartimento di Informatica

Hate Speech Monitoring Group

HATE SPEECH AND SOCIAL MEDIA

I ♥ HATE PREJUDICE

contro l'odio

Linguistica Computazionale

(o Elaborazione del Linguaggio Naturale)

Processamento automatico del linguaggio tramite applicazioni dell'Informatica.

- Algoritmi
- Risorse linguistiche
- Grandi quantità di dati

Linguistica Computazionale

Applicazioni

- Traduzione automatica
- Assistenti virtuali
- Classificazione (es. filtro spam)
- Analisi del sentimento
- ...

Linguistica Computazionale

Aree

- Morfologia → forma delle parole
- Fonetica → suono delle parole
- Sintassi → relazione tra parole
- Semantica → significato delle parole
- Pragmatica → intenzione della comunicazione
- …
- Apprendimento automatico
- Statistica

Università di Groningen (NL), 2012

PhD in Semantica Computazionale

Università di Groningen (NL), 2012

Raccolta di messaggi Twitter in lingua italiana.

100 milioni di tweet in un anno.

Pietro Salvatori
@PietroSalvatori

Follow

"@HuffPostItalia: #Grillo arrivato nel backstage huff.to/1afz4P8 #m5s"

9:34 PM - 24 May 2013

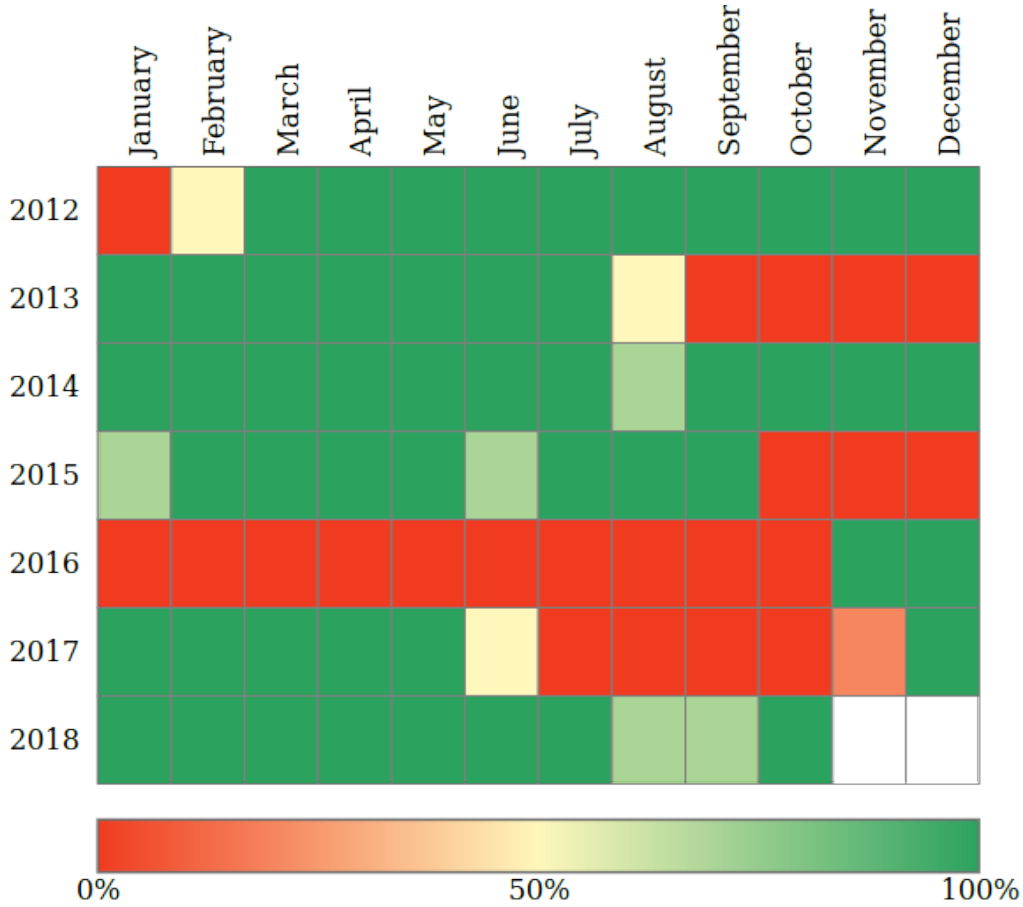
Query: `vita Roma forza alla quanto amore Milano Italia fare grazie della anche periodo bene scuola dopo tutto ancora tutti fatto`

twita

http://twita.di.unito.it/

~1 miliardo di tweet
2012-2020

Metodo di raccolta
aggiornato



Valerio Basile, Mirko Lai, Manuela Sanguinetti: *Long-term Social Media Data Collection at the University of Turin,* CliC-it 2018

Università di Groningen (NL), 2012

Killer app: Sentiment Analysis
(o Opinion Mining)

Estrazione automatica del sentimento
(o polarità) espresso in un messaggio
in linguaggio naturale

Sentiment Analysis

- **Supervisionato**
  apprendimento automatico a partire da corpora annotati

- **Non supervisionato**
  basato su risorse come lessici affettivi

Lessico affettivo

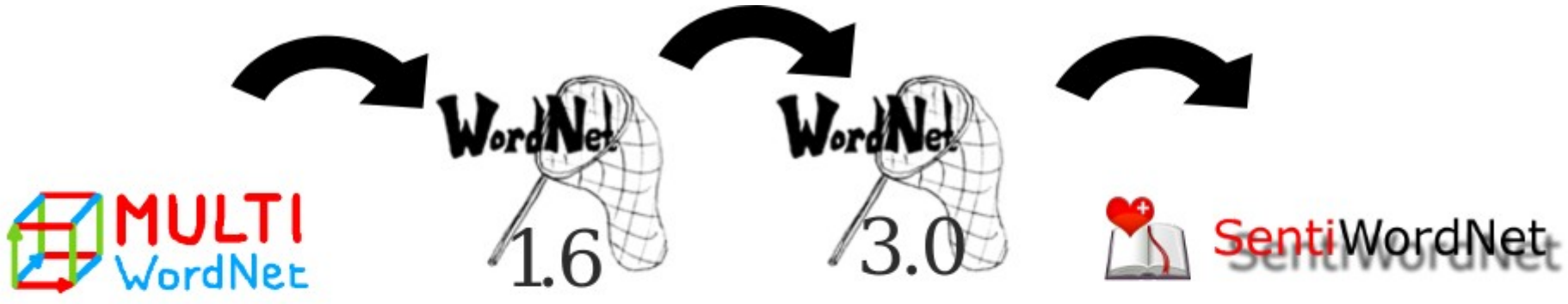
Elenco di parole (forme o lemmi) associate ad uno o più punteggi numerici o categoriali.

es.

```
Terribile (agg.)      -0.80
Decentemente (avv.)    0.25
Fantastico (agg)       0.70
Arrabbiarsi (v.)      -0.50
```

Lessico affettivo

Costruzione di un lessico affettivo per l'italiano
Allineamento di risorse linguistiche (WordNet)



invidia (n) 00758335
invidia (n) 07549716

00758335 (n) envy, invidia (spite and resentment at seeing the success of another (personified as one of the deadly sins))
07549716 (n) envy, enviousness (a feeling of grudging admiration and desire to have something that is possessed by another)

n#00758335 0 0.5
n#07549716 0.625 0

Lessico affettivo

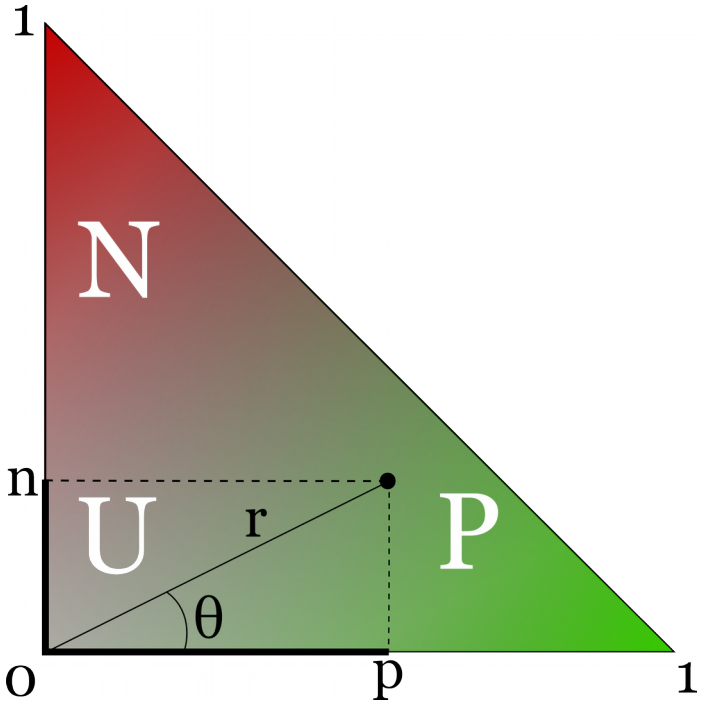
Sentix

41K lemmi
74K sensi
Nomi, Verbi, Aggettivi, Avverbi
Punteggi di polarità e intensità

Valerio Basile, Malvina Nissim:
Sentiment analysis on Italian tweets, WASSA 2013

Lessico affettivo

lemma	POS	Wordnet synset ID	positive score	negative score	polarity	intensity
naturale	A	00074346	0,75	0,125	0,789	0,76



$$\theta = tan^{-1}(\frac{n}{p})$$

$$r = \sqrt{p^2 + n^2}$$

$$1 - \frac{4\theta}{\pi}$$

$$\iota = r$$

Sentiment Analysis

Algoritmo basilare:

- Pipeline NLP
  testo → token → lemma → polarità
- Somma (o media) dei valori di polarità
- Valori soglia per categorie
  positivo/neutro/negativo

Test addizionali: part-of-speech, polipatia

Lessico affettivo

Valutazione basata su 2,000 tweet annotati manualmente:

- 1,000 tweet generici

	positive	negative	neutral
best precision	0.440 (r)	0.195 (v)	0.664 (nar)
best recall	0.701 (nva)	0.532 (var)	0.669 (a)
best F-score	0.485 (nvar)	0.262 (vr)	0.647 (a)

- 1,000 tweet su argomenti politici

	positive	negative	neutral
best precision	0.164 (a)	0.412 (a)	0.617 (nar)
best recall	0.593 (nva)	0.150 (nr)	0.724 (a)
best f-score	0.251 (nv)	0.213 (nr)	0.637 (a)

Lessico affettivo

Sentix 2.0

- Pulitura automatica
- Solo lemmi
  → sensi collassati con media pesata
- Valore unico → polarità * intensità

Lessico affettivo

Altri lessici affettivi per l'italiano

- OpeNER Sentiment lexicon
- Proprietary sentiment lexicon (CELI)
- Polarized word embeddings by G. Attardi (IIR 2015)
- Norme affettive (INCREASE)
- Automatic method to build multilingual opinionated lexicons, G. Castellucci, D. Croce, R. Basili (2016)
- SentiWords (FBK)
- SentIta and Doxa: Italian databases and tools for sentiment analysis.
- Sviluppati per EVALITA: UNIBA; Di Gennaro, Rossi e Tamburini

http://www.ai-lc.it/lessici-affettivi-per-litaliano/

CREA, Roma, 2018

Applicazione di Sentix per l'analisi del sentimento nel dominio agroalimentare.

L'analisi statistica fa emergere problemi nell'approccio dictionary-based.

Propagazione dell'errore nella pipeline testo → token → lemma → polarità

CREA, Roma, 2018

Test di correttezza con dizionario HOEPLI

Original	@ANBI_Nazionale Allarme idrico. Dopo il Po anche l'Adige è in crisi d'acqua https://t.co/GLTlMNqzEv di @AgriculturaIT
ISDT	acqua **adigire** allarme crisi **d** dopo idrico po - Sentix score: 0.080
POSTWITA	acqua **adigere** allarme crisi di dopo idrico po - Sentix score: 0.080
PARTUT	**acquare adigere** allarme crisi **d** dopo idrico po - Sentix score: -0.078

Original text	Capitale Europea della Cultura che combacia con la fine consultazioni de #labuonascuola: gran bel segnale :)
Bag of words	bel Capitale combacia consultazioni Cultura della Europea fine gran segnale
ISDT	bello capitale combaciare consultazione cultura di europeo fine grande segnale - Sentix score: 0,8449
POSTWITA	bello capitale combaciare consultazione cultura **da** europeo fine grande segnale - Sentix score: 1,0739
PARTUT	**bel** capitale **combacia** consultazione cultura **dere** europeo fine grande segnale - Sentix score: -0,2715

CREA, Roma, 2018

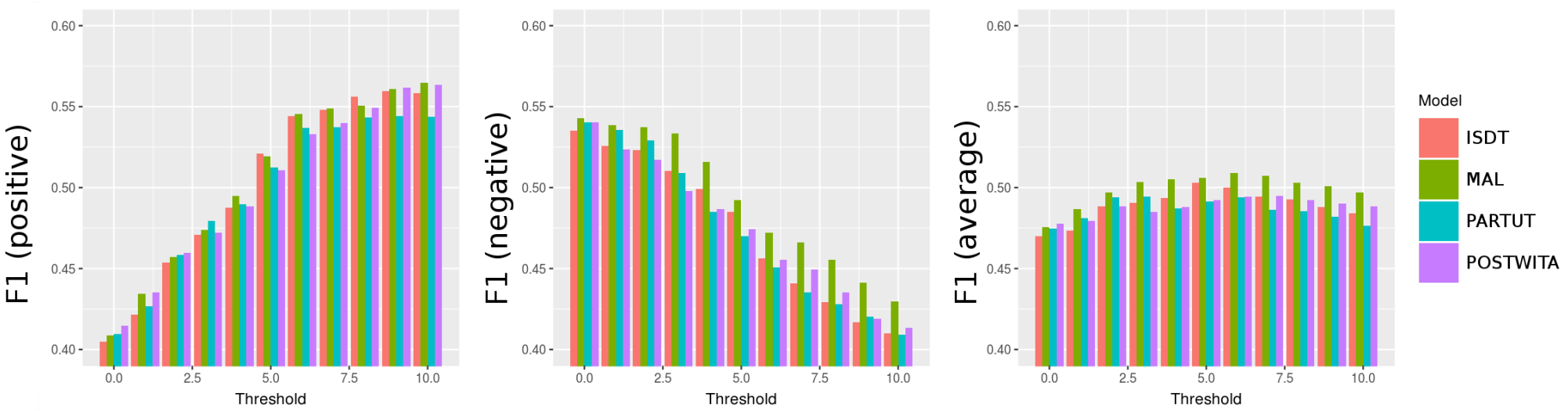
MAL
(Morphologically-inflected Affective Lexicon)

Espansione di Sentix dai lemmi alle forme usando *Morph-it!*

Migliore performance di analisi del sentimento su SENTIPOLC e sul nuovo dataset AGRItrend

CREA, Roma, 2018

Analisi dello sbilanciamento di polarità nei lessici → work in progress



Marco Vassallo, Giuliano Gabrieli, Valerio Basile and Cristina Bosco: *The Tenuousness of Lemmatization in Lexicon-based Sentiment Analysis* (CliC-it 2019)

Campagne di valutazione

EVALITA

Campagna di valutazione per le tecnologie del linguaggio italiano

http://www.evalita.it/

Organizzato da

AILC

Associazione Italiana di Linguistica Computazionale

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

Campagne di valutazione

- Prima edizione: 2007
- 7ª edizione: 2020
- Organizzata in shared task
  2014 → 7 task, 2016 → 6 task, 2018 → 10 task, 2020?
- Sentiment Analysis inizia nel 2014

SENTIPOLC

SENTIment POLarity Classification

- Task più partecipato a EVALITA 2014
- 4513+1935 tweet annotati
- SentiTUT (Torino), TWITA (Groningen)
- Tweet generici e tweet politici
- Soggettività, Polarità, Ironia

SENTIPOLC

subj	pos	neg	iro	description
0	0	0	0	an objective tweet example: *l'articolo di Roberto Ciccarelli dal manifesto di oggi* `http://fb.me/1BQVy5WAk`
1	0	0	0	a subjective tweet with neutral polarity and no irony example: *Primo passaggio alla #strabrollo ma secondo me non era un iscritto*
1	1	0	0	a subjective tweet with positive polarity and no irony example: *splendida foto di Fabrizio, pluri cliccata nei siti internazionali di Photo Natura* `http://t.co/GWoZqbxAuS`
1	0	1	0	a subjective tweet with negative polarity and no irony example: *Monti, ripensaci: l'inutile Torino-Lione inguaia l'Italia: Tav, appello a Mario Monti da Mercalli, Cicconi, Pont...* `http://t.co/3CazKS7Y`
1	1	1	0	a subjective tweet with positive and negative polarity (mixed polarity) and no irony example: *Dati negativi da Confindustria che spera nel nuovo governo Monti. Castiglione: "Avanti con le riforme"* `http://t.co/kIKnbFY7`
1	1	0	1	a subjective tweet with positive polarity, and an ironic twist example: *Letta: sicuramente non farò parte del governo Monti . e siamo un passo avanti. #finecorsa*
1	0	1	1	a subjective tweet with negative polarity, and an ironic twist example: *Botta di ottimismo a #lInfedele: Governo Monti, o la va o la spacca.*

SENTIPOLC

- Benchmark per SA sull'italiano
- Superiorità dei sistemi supervisionati (reti neurali ricorrenti e convoluzionali, SVM)
- Seconda edizione nel 2016

V. Basile, A. Bolioli, M. Nissim, V. Patti and P. Rosso: *Overview of the Evalita 2014 SENTIment POLarity Classification Task* (EVALITA'14)

V. Basile, F. Barbieri, D. Croce, M. Nissim, N. Novielli, V. Patti: *Overview of the EVALITA 2016 SENTiment POLarity Classification Task* (EVALITA'16)

Oltre il sentiment

- Aspect-Based Sentiment Analysis
  Primo task a EVALITA 2018
  (ABSITA, Unito+Uniba)

- Rilevazione dell'ironia e del sarcasmo
  Corpus TWITTIRò (Alessandra Cignarella)
  IronITA @ EVALITA 2018

Oltre il lessico

I sistemi supervisionati forniscono predizioni più accurate. Caveat:

- Poca trasparenza
- Necessità di grandi quantità di dati
- Complessi e computazionalmente costosi

Oltre il lessico

AlBERTo

Modello per l'italiano della rete neurale basata su transformer network BERT (Google)

Stato dell'arte su SENTIPOLC e Hate Speech

M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile: *AlBERTo: Italian BERT Language Understandin Model for NLP Challenging Tasks Based on Tweets,* CliC-it 2019

https://github.com/marcopoli/AlBERTo-it

Grazie!

al CREA
a Marco & Giuliano

http://valeriobasile.github.io/