

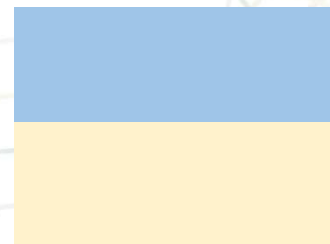
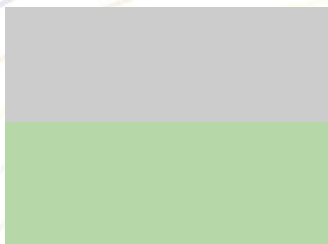
Sentiment Analysis con il CREA

Tutto parte da Sentix

Valerio Basile
Università di Torino

Nota biografica

- PhD a Groningen
- Postdoc Inria Sophia Antipolis
- Ricercatore Università di Torino



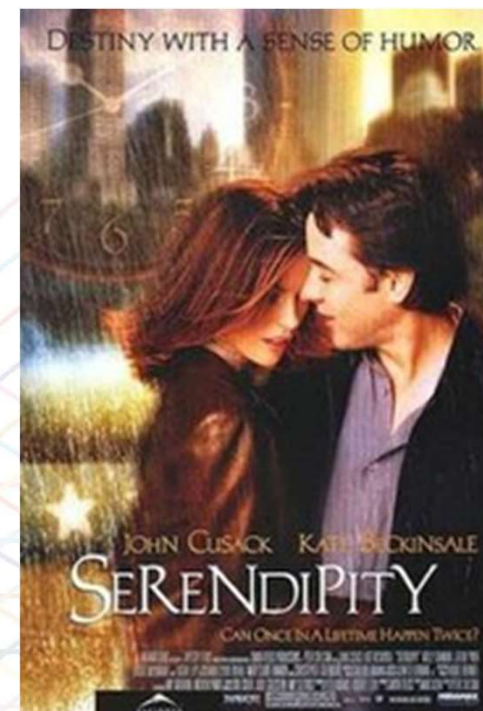
**UNIVERSITÀ
DI TORINO**

"Serendipity"

Anno 2012

Hot topics:
Social Media (Twitter)
Sentiment Analysis

Si comincia con la raccolta dati...



Collezione di **tweet** in
lingua italiana in
modalità "streaming"

Prima strategia:
lemmi frequenti
-> lang. id.

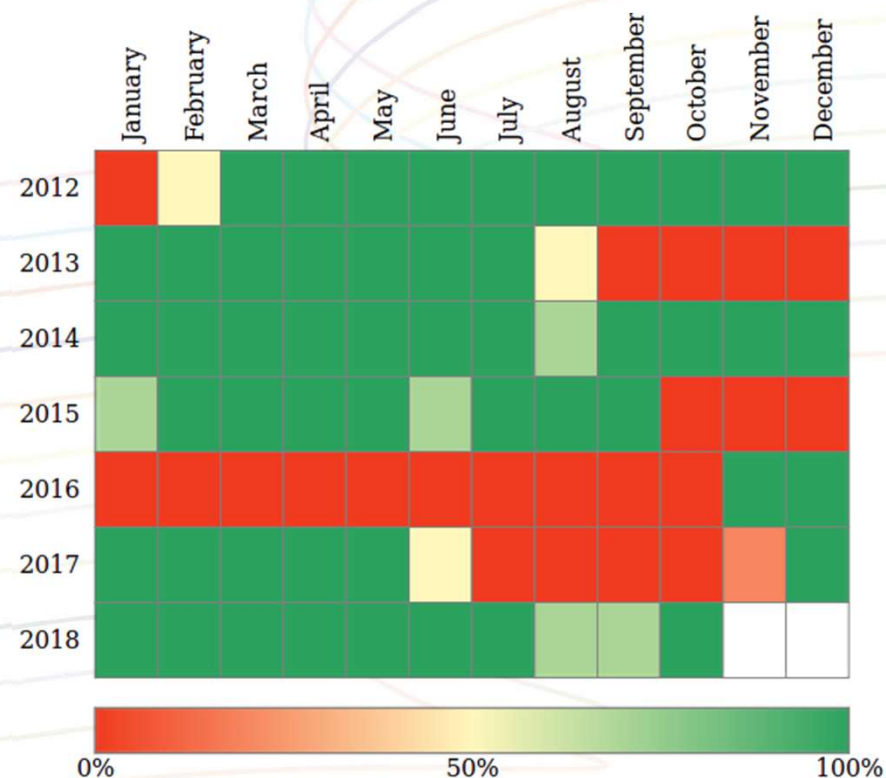
Dataset "traslocato"
da Groningen a Torino

twito



 Basile, Lai, Sanguinetti (2018)
*Long-term Social Media Data Collection at
the University of Turin*
CLiC-it 2018

500 milioni di tweets!
(1B oggi?)



In Natural Language Processing, spesso non basta **pubblicare** una risorsa...

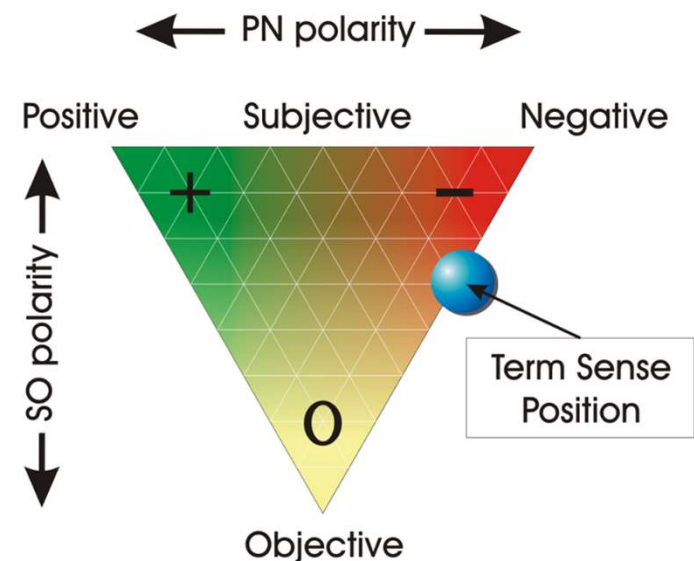
 Basile e Nissim (2013)
Sentiment Analysis on Italian Tweets
WASSA 2013



Allineamento tra risorse

SentiWordNet
(Esuli e Sebastiani 2006)

Sensi di parole
classificati
automaticamente
in base alla glossa



$P = 0.75, N = 0, O = 0.25$	<u>estimable(1)</u> <i>deserving of respect or high regard</i>
$P = 0.625, N = 0.25, O = 0.125$	<u>honorable(5)</u> <u>good(4)</u> <u>respectable(2)</u> <u>estimable(2)</u> <i>deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"</i>
$P = 0, N = 0, O = 1$	<u>computable(1)</u> <u>estimable(3)</u> <i>may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"</i>

WordNet

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Adjective

- [S:](#) (adj) **estimable** (deserving of respect or high regard)
- [S:](#) (adj) **estimable**, [good](#), [honorable](#), [respectable](#) (deserving of esteem and respect)
"all respectable companies give guarantees"; "ruined the family's good name"
 - [similar to](#)
 - [derivationally related form](#)
 - [antonym](#)
 - [W:](#) (adj) [disreputable](#) [Indirect via [reputable](#)] (lacking respectability in character or behavior or appearance)
- [S:](#) (adj) [computable](#), **estimable** (may be computed or estimated) *"a calculable risk"; "computable odds"; "estimable assets"*

Dizionario elettronico
Lemmi N-to-N Sensi

MultiWordNet

 Pianta, Bentivogli, Girardi
*MultiWordNet: developing an
aligned multilingual database*
Conference on Global WordNet
2002

Sense: **agricoltura/1**

Pos: Noun

Domain: Animal_Husbandry,
Enterprise

Affective:



Synset: **farming, agriculture,
husbandry**

Phraset:

Gloss: the practice of cultivating
the land or raising stock



Synset: **agraria, agricoltura**

Phraset:

Gloss:

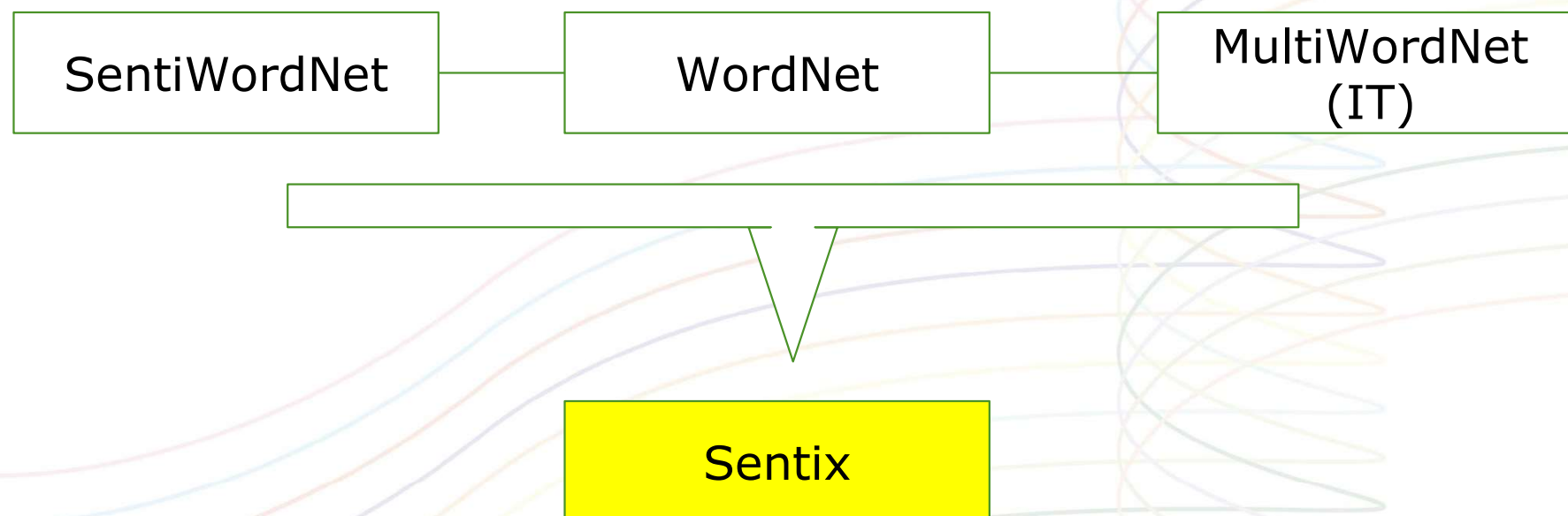


Synset: **agricultura, explotación,
granjería, labranza**

Phraset:

Gloss:

Allineamento tra risorse esistenti



Sentix

Basile e Nissim (2013)

Sensi con punteggi di polarità
(positivo/negativo)

Sentix 2.0

41800 item

Sensi -> Lemmi

Pos/neg -> Polarità+intensità $[-1, 1]$

Sentix

Basile e Nissim (2013)

```
irrefutabilità 0.11327372626976256
contrapposizione 0.0
stramazzare -0.25
organismi_eucarioti 0.375
mortale -0.625
trochilidae 0.375
triangolo_acuto 0.25
posse 0.125
durevoli_di_consumo 0.11450499647135175
epilettica 0.25
```

Sentix è divenuto molto popolare,
soprattutto in comunità fuori da CL/NLP
(statistica, text analysis, social sciences)

Strumento semplice e disponibile...

...ma **perfettibile**:
copertura, rumore, bias

Nel 2017-2018 comincia la collaborazione
con il CREA (Roma)

L'applicazione di Sentiment Analysis a dati
di social media su temi dell'agricoltura
porta **inaspettati** risultati.

Vassallo, Gabrieli, Basile, Bosco *The Tenuousness of Lemmatization in Lexicon-based Sentiment Analysis* CLiC-it 2019

SA con Sentix su SENTIPOLC+AGRITREND

Lemmatizzazione:

Original	@ANBI.Nazionale Allarme idrico. Dopo il Po anche l'Adige è in crisi d'acqua https://t.co/GLTlMNqzEv di @AgricolturaIT
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

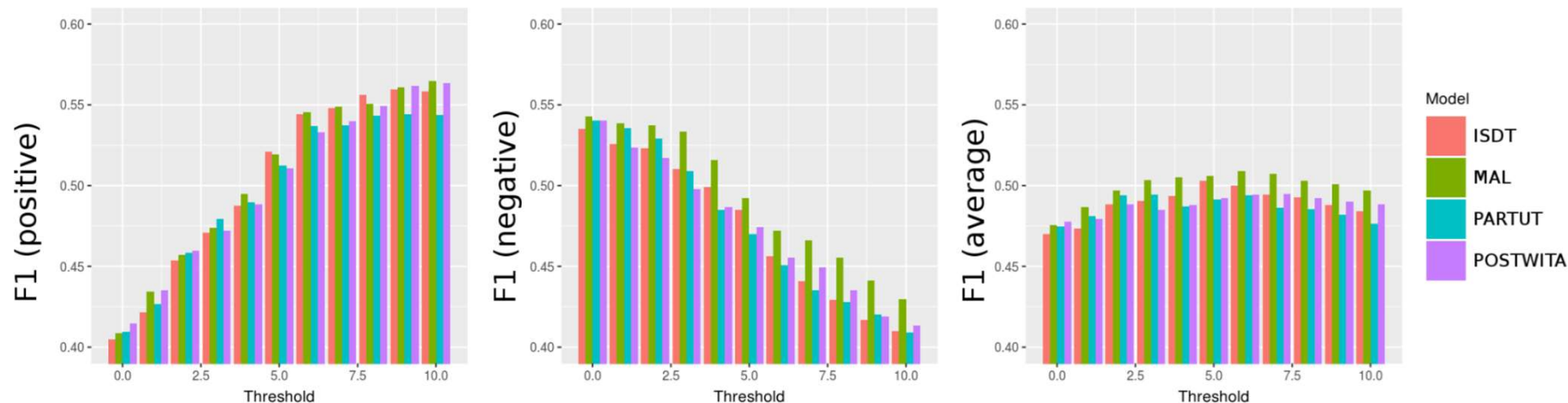
E se potessimo saltare il passaggio della
lemmatizzazione?

MAL
Morphologically-inflected Affective Lexicon

Tramite l'analizzatore morfologico morph-
it (Zanchetta e Baroni, 2005)
espandiamo i lemmi di Sentix
in tutte le loro **forme**

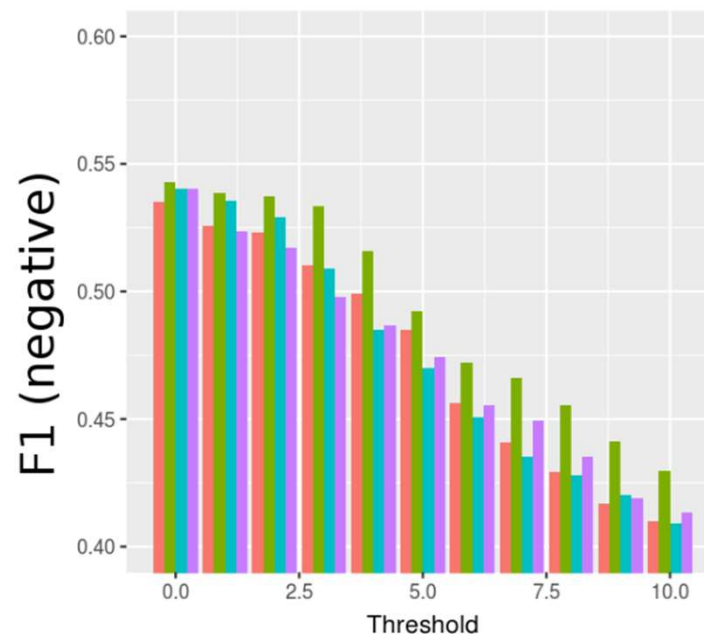
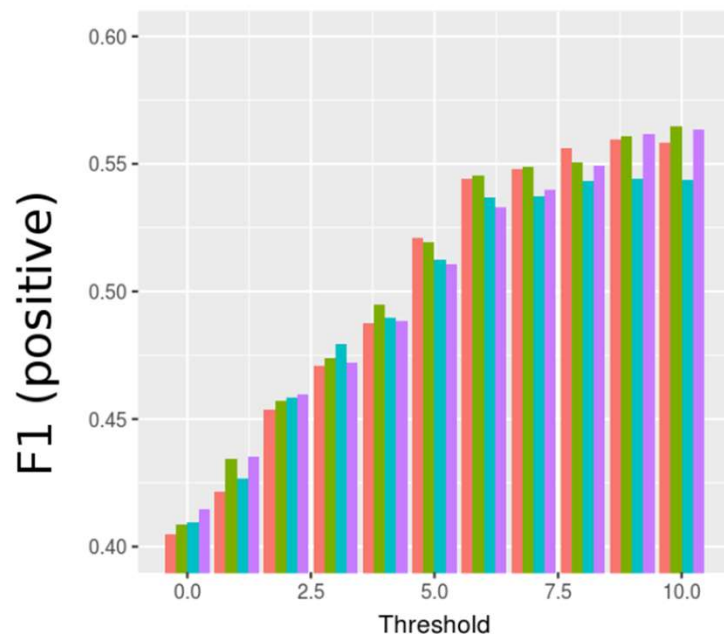
-> 148,867 forme distinte
+ controllo con dizionario Hoepli





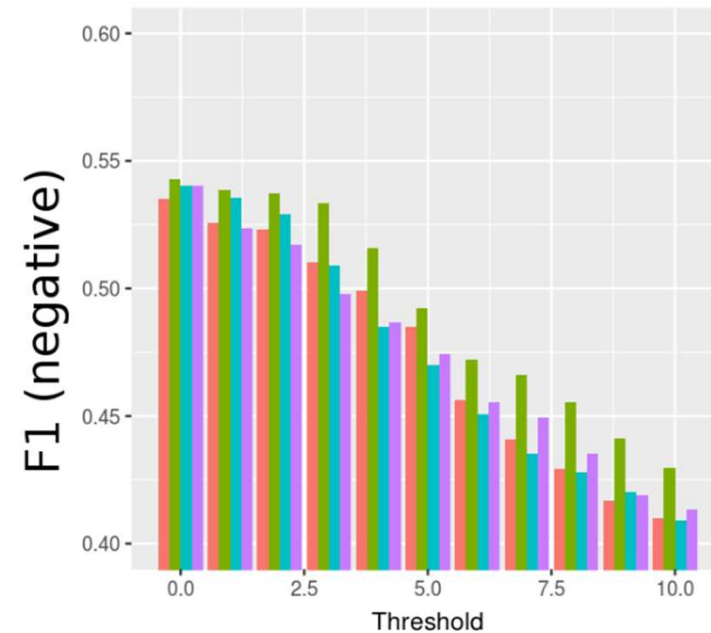
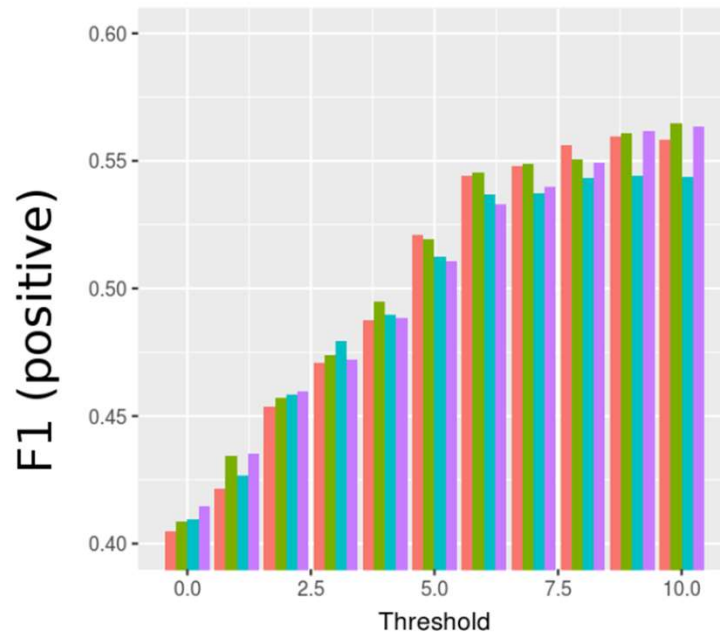
MAL conduce una SA migliore di Sentix in particolare nella predizione del sentimento **negativo**

Asse X = valore soglia per decidere la label



Guardiamo questi grafici con più attenzione...

La sensibilità alla soglia è diversa tra positivo vs. negativo



 Vassallo, Gabrieli, Basile, Bosco
*Polarity Imbalance in Lexicon-based
Sentiment Analysis*, CLiC-it 2020

Analisi sistematica del fenomeno di
sbilanciamento

Osservazione: l'argomento (topic)
influenza lo sbilanciamento:
generale -> meno sbilanciamento
politico -> più sbilanciamento

Alcuni topic presentano bias di polarità
(es. "guerra" - negativo)

Domanda di ricerca: lo sbilanciamento
viene dal **lessico**, dai **dati**, o da entrambi?

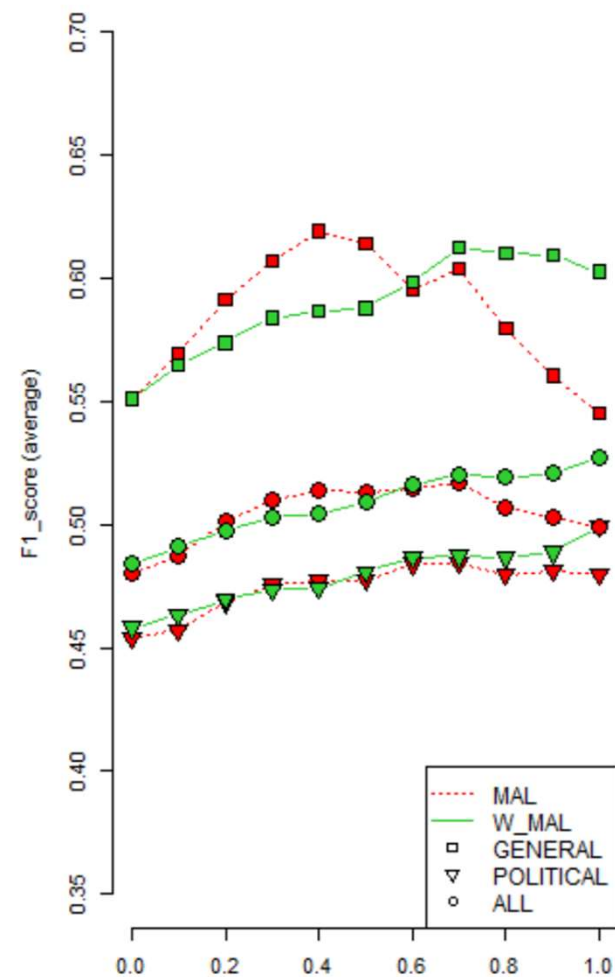
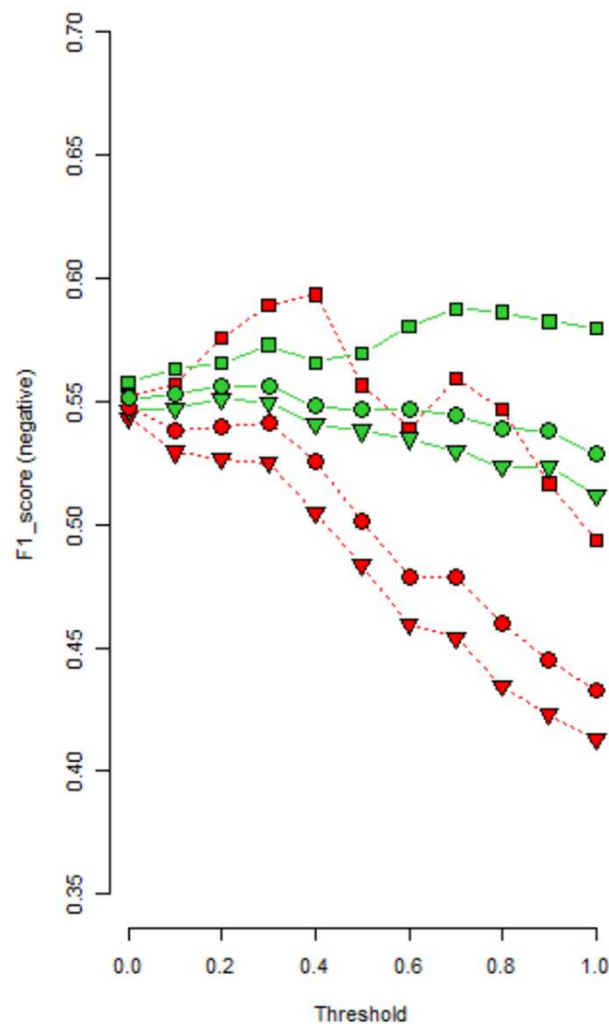
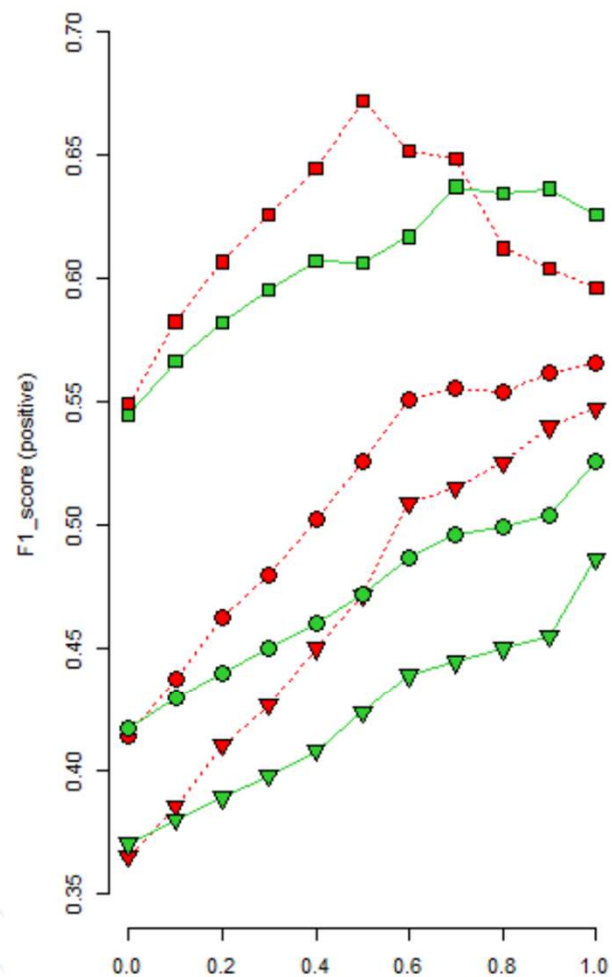
Idea: “pesare” i termini del MAL con l’informazione sulla frequenza delle parole

$$Zipf(i) = \log_{10} \left(\frac{f(i)}{\frac{\sum_{i=1}^N f(i)}{10^6} + \frac{N}{10^6}} \right) + 3$$

Frequenze calcolate su TWITA

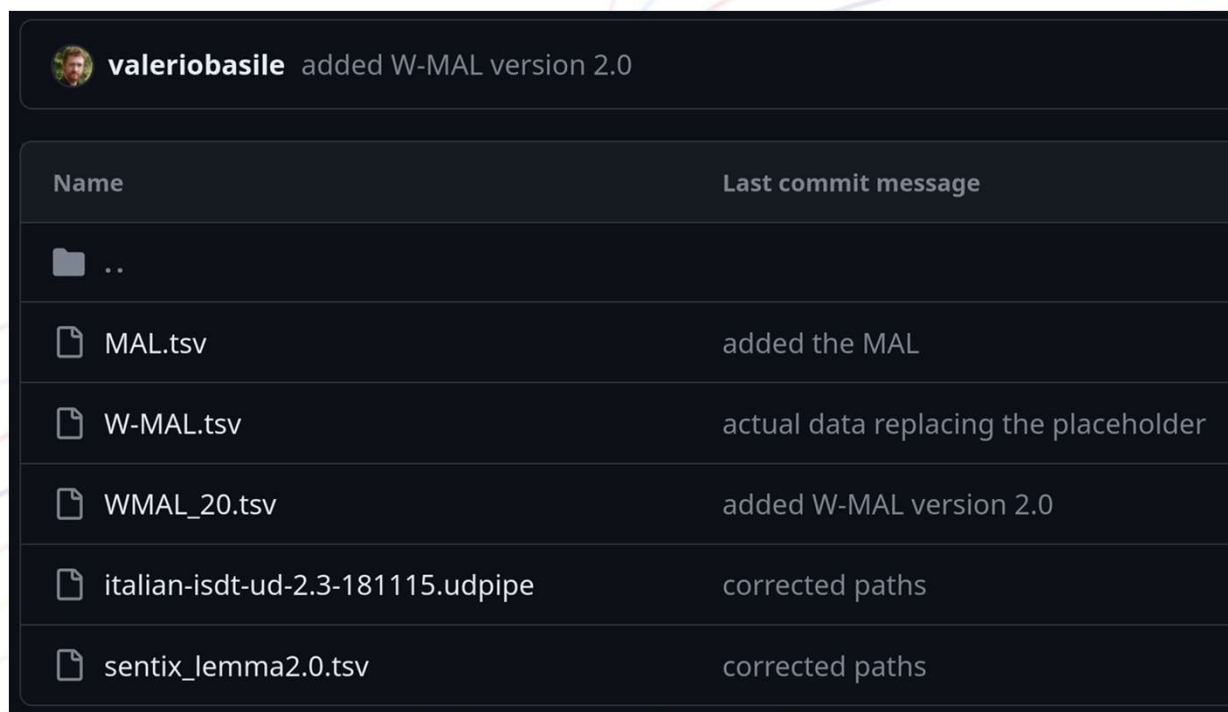
Intuizione:

“more sought-after terms are often used to convey stronger opinions and feelings”



Sentix, MAL, W-MAL

Stato dell'arte tra lessici affettivi per la lingua italiana



A screenshot of a GitHub commit interface. At the top, a user profile icon and the name 'valeriobasile' are shown next to the text 'added W-MAL version 2.0'. Below this is a table with two columns: 'Name' and 'Last commit message'. The table lists several files and folders with their corresponding commit messages.

Name	Last commit message
..	
MAL.tsv	added the MAL
W-MAL.tsv	actual data replacing the placeholder
WMAL_20.tsv	added W-MAL version 2.0
italian-isdt-ud-2.3-181115.udpipe	corrected paths
sentix_lemma2.0.tsv	corrected paths

<https://github.com/valeriobasile/sentixR>

Sentix, MAL, W-MAL

Usati, tra le altre applicazioni, per CreAgritrend



1. IL TERMOMETRO DELL' AGRICOLTURA

I PRINCIPALI TEMI DISCUSSI SU TWITTER E GLI UMORI DEGLI ADDETTI

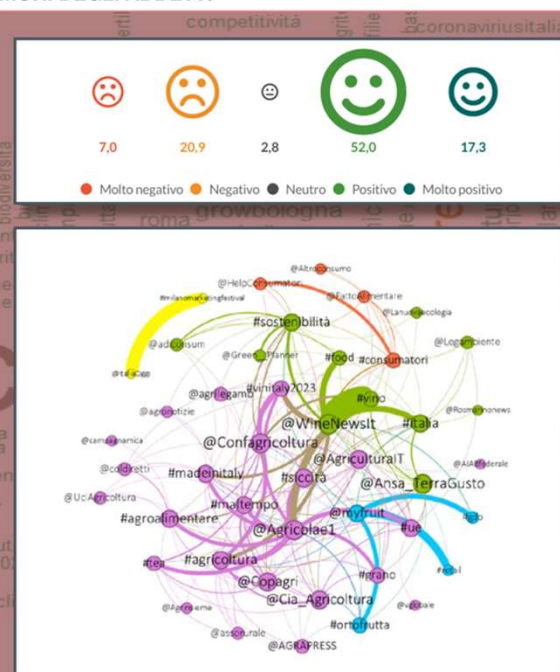
L'analisi del **sentimento** del settore agricolo e agroalimentare ha focalizzato l'attenzione su 13.138 tweet postati da media specializzati (85,4% dei tweet raccolti) e parti sociali (restante 14,6%) del settore agroalimentare italiano, dal 9 marzo all' 11 giugno 2023. Questo periodo è stato caratterizzato da una serie di eventi, tra cui una eccezionale ondata di maltempo, che tra il 16 e il 17 maggio scorso si è abbattuta principalmente sull'Emilia-Romagna. L'enorme massa d'acqua caduta in poche ore ha causato l'esondazione di diversi fiumi e ha messo in ginocchio sia la popolazione che l'intero settore agricolo della regione.

Il sentimento generale nel settore è rimasto comunque di fiducia, con il 69,3% di giudizi positivi che confermano la tendenza registrata anche nel trimestre precedente. Sia i giudizi negativi pari a 27,9%, che quelli neutrali (2,8%) hanno subito una lieve diminuzione rispetto al trimestre precedente, rispettivamente di -0,8% e -1,4%. Probabilmente a rafforzare il sentimento di fiducia ha contribuito anche il risultato delle elezioni amministrative svoltesi a metà maggio, che hanno confermato l'attuale governo, garantendo una certa stabilità istituzionale.

L'analisi delle reti sociali (Social Network Analysis), chiamata rete bimodale con partizione modulare, applicata alle tematiche (#hashtag) maggiormente utilizzate dagli utenti (@user) (dove lo spessore degli archi in figura rappresenta il grado (frequenza) che unisce l'utilizzo di un hashtag da parte di quell'utente), ha individuato cinque gruppi omogenei.

Il primo gruppo (viola) è quello più numeroso e raggruppa una serie di tematiche legate sia agli eventi climatici che alla promozione del Made in Italy delle produzioni agroalimentari. Il periodo in esame, infatti, è stato caratterizzato inizialmente da una forte siccità a cui però è seguito un lungo intervallo caratterizzato da forti precipitazioni che, come detto in precedenza, hanno provocato ingenti danni all'intero settore.

Il secondo gruppo (verde) pone l'accento sul settore vitivinicolo



Ma soprattutto, eccellente **collaborazione**
nazionale interdisciplinare

Accordo quadro UniTO-CREA
Workshops, Partecipazioni a conferenze

Nel 2020 l'asse diventa un triangolo con **Bari**

