A proposal for an integrated approach between sentiment analysis and social network analysis

Domenico De Stefano[1] and Francesco Santelli [1]

[1]Department of Political and Social Sciences,
University of Trieste

ddestefano@units.it

fsantelli@units.it

StaTalk2019 – 2019 - November 22, Trieste

Background

- In Online Social Media Data (OSMD) key role played by Twitter, especially for the analysis of opinions spreading (Go et al., 2009; Onorati, Diaz, 2016).
- Twitter is a microblogging service where users tweet about any topic within the 140-character limit and follow others to receive their tweets. Usually tweets are tagged with #hashtags

In Twitter studies:

- focus is given to the analysis of the sentiment about a given topic OR to the analysis of the social network among users or tweets $\Rightarrow$ rarely both approaches are combined
- concern Retweets analysis to understand the mechanism and dynamics of opinion flow (Suh at al., 2010;Rossi, Magnani, 2012)
- Other kind of interactions, i.e. mentioning a user or reply to a tweet are usually not considered

Aim of the study

Determine the structural characteristics of opinion diffusion about a topic on Twitter

Reconstruct not only the tweet-retweet but also the tweet-reply chains of opinions about a trending topic on Twitter implementing some additional elements related to the semantic field of tweets

- Message-based perspective; not user-based perspective!
- Multi-steps procedure to derive a signed network, related to the structure of spread of contents and opinions
- Sentiment analysis algorithms determine the sign of each link and the structure of the obtained network gives insights on how opinion diffuse on the platform

Our procedure: details

- We start from a population of tweets on a given reference topic
  - Original tweets $\rightarrow$ tweets typed by users containing the original topics as hashtag
  - Retweets $\rightarrow$ Fixed content no further text is added
  - Replies $\rightarrow$ Answers to the original tweet, users leave a comment
- To analyze the opinion spreading we adopt a procedure consisting in:
  1. Reducing tweets dimensionality $\Rightarrow$ Extracting concepts
  2. First step sentiment analysis $\Rightarrow$ Groups of tweets polarity
  3. (Conditional) sentiment analysis $\Rightarrow$ Concepts sentiment spreading
  4. Analyze the obtained signed networks (for each concepts)

Findings are, then, related to both individual field of communication (semantic) and communal activities (network). This is consistent with communication studies such as Murthy, 2013.

Step 1: Reducing tweets dimensionality /1

- The *i*-th tweet, $i = 1, \ldots, n$ is marked by a number of Hashtags expressing, in few words, the subject of the tweet wrt the reference topic.
- Hashtags that occur frequently together in same tweets describe a common latent structure.
- Assuming m the total number of hashtags a $m \times m$ matrix can be defined

$$ADJ = \begin{array}{c} \\ salvini \\ 80euro \\ lega \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} salvini & 80euro & lega & \ldots \end{array} \\ \begin{pmatrix} . & 3 & 7 & \ldots \\ 3 & . & 3 & \ldots \\ 7 & 3 & . & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

A small subset of the hashtag co-occurrences matrix.

- Fast greedy algorithm, suited for symmetric, weighted and undirected graph.
- It finds communities minimizing a quantity called modularity $\rightarrow$ $Q = \sum_u \left\{ \frac{e_{uu}}{2m} - \left(\frac{a_u}{2m} \right)^2 \right\}$. Each u is a cluster, $\frac{a_u}{2m}$ is expected fraction of edges in u, , e_{uu} is the number of edges connecting vertices in cluster u.

Step 1: Reducing tweets dimensionality /2

- Each community of hashtags thus identified expresses a concept. Next step: assign tweets to the concept u.

$$TweetsConcepts = \begin{array}{c} \\ Tweet1 \\ Tweet2 \\ Tweet3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} Concept1 & Concept2 & Concept3 & ... \end{array} \\ \begin{pmatrix} 2 & 1 & 0 & \dots \\ 0 & 0 & 3 & \dots \\ 0 & 2 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

A subset of the tweets-concepts matrix.

- Each tweet can include hashtags related to several concepts.
- Tweets classification based on both automatic hierarchical clustering but also qualitative evaluation (communities are a strong hint but some qualitative analysis are worthy).
- From thousands of original tweets to few groups of tweets related, semantically, to latent concepts.

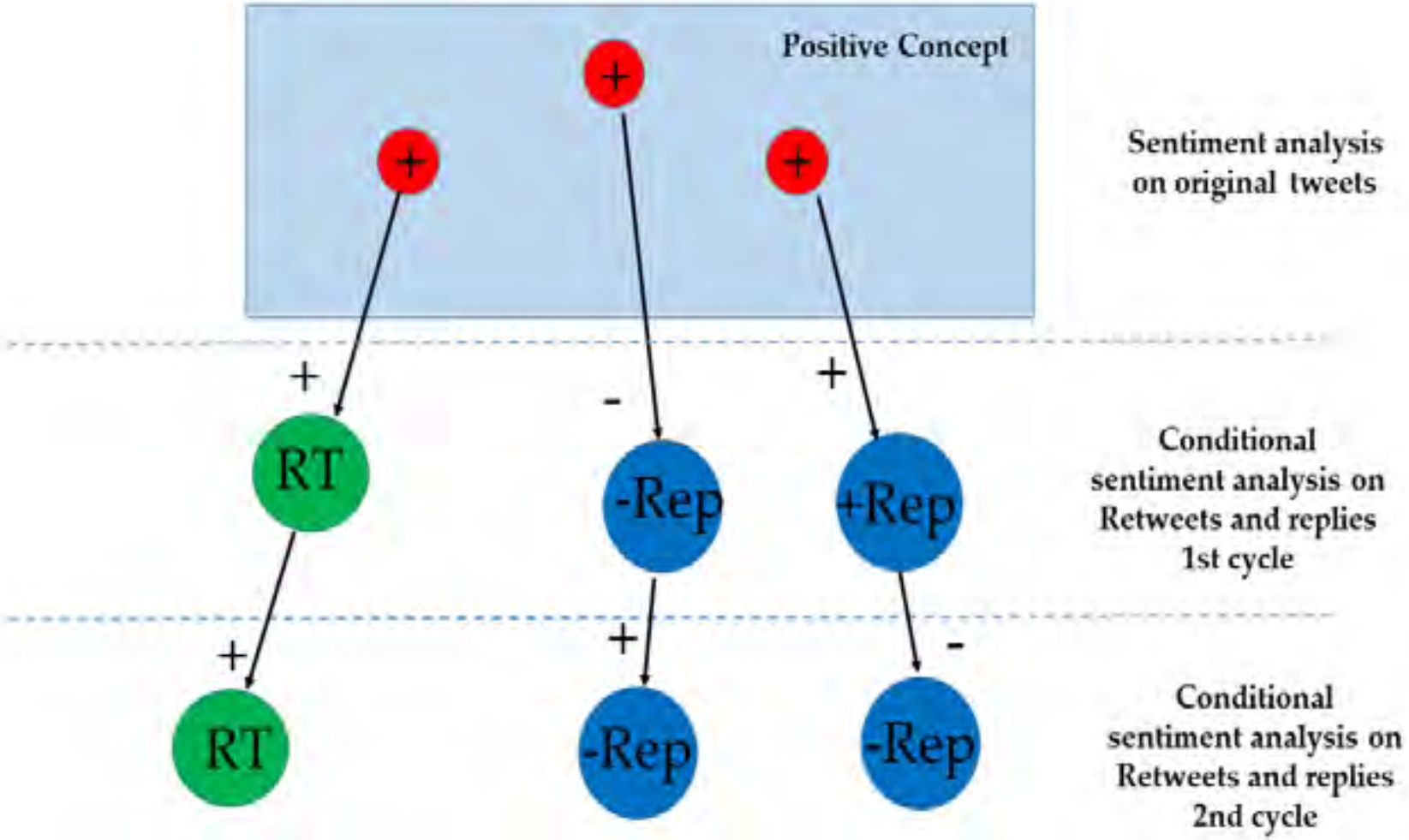
Step 2: First step sentiment analysis

- After the dimensionality reduction step we use Sentiment analysis to determine:
- The sign of the original tweets (wrt the reference topic).
  Main ideas behind it:
  1. The sentiment is related to the concept (based on hashtags).
  2. From tweets to few concepts $\Rightarrow$ Sentiment is based on both automatic and manual classification of hashtags within concepts.
  3. Hashtags have been used firstly as neutral entities, marking tweets by their reference topic. Here, few of them they are used also as positive/negative entities.
  4. General assumption: the procedure can not be completely automatic in each step, some human interventions by the researcher are required to improve quality of procedure.

Step 3: (Conditional) sentiment analysis /1

- After the first step Sentiment Analysis (original tweet), next phase, to study opinion spreading based on sentiment, is the Sentiment Analysis on retweet-reply chains.
- The sign of the original tweets (wrt the reference topic) is now given $\Rightarrow$ the sign of the edges connecting the original tweet to the retweets and to the replies (conditional to the concept).
- Replies may or may not have hashtags $\Rightarrow$ Stemming is necessary
- Retweets have the same sentiment of the attached concept; usually no further text is added.
- Replies sentiment depend upon the concept hashtags

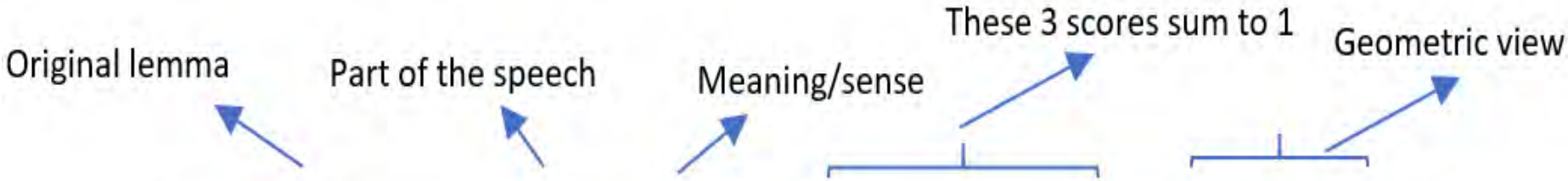
Step 3: (Conditional) sentiment analysis /2



Our approach to Sentiment Analysis /1

- Among several kinds of sentiment analysis approaches, we have adopted so far a procedure that links tweets lemma to a polarity lexicon of Italian language, no matter what is the context (Basile, Nissim, 2013).

Lexicon structure



	A	B	C	D	E	F	G	H
1	***lemma***	***POS***	***synset_ID***	***positive***	***negative***	***neutral***	***polarity***	***intensity***
2	abile	a	00001740	0,13	0,00	0,88	1,00	0,13
3	intelligente	a	00001740	0,13	0,00	0,88	1,00	0,13
4	valente	a	00001740	0,13	0,00	0,88	1,00	0,13
5	capace	a	00001740	0,13	0,00	0,88	1,00	0,13
6	incapace	a	00002098	0,00	0,75	0,25	-1,00	0,75
7	assoluto	a	00005205	0,50	0,00	0,50	1,00	0,50
8	relativo	a	00006032	0,25	0,50	0,25	-0,41	0,56
9	astinente	a	00009346	0,00	0,63	0,38	-1,00	0,63

Our approach to Sentiment Analysis /2

- Original tweets have been preprocessed in order to be analyzed: stemming, removing stopwords, removing punctuation and so on.
- Then, to create a potential join between lexicon spreadsheet and tweets data, text has been tagged using Treetagger (Schmid, 1994; Baroni, 2005).
- Thus, each tweet is now expressed by lemmas and, after the join, by lemmas 5 scores: positive, neutral, negative, polarity and intensity.

Issues in automatic join

- Some lemmas from Treetagger are not the same in Lexicon (miss-join).
- Some lemmas has more than 1 meaning (synsets). Solution in automatic procedure:
  1. Average scores across synsets
  2. Remove lemmas with high standard deviation in polarity scores (too ambiguous). We have decided to remove 25% most ambiguous terms.

Real data case: #flattax

- A flat tax system applies the same tax rate to every citizen regardless of their income bracket. It is a leitmotif of Northern League (Lega Nord) party.
- The system is of course not easily applicable due to its economic cost for public expenditure.
- In the last Economic and Financial Document of the 9^{th} April, 2019, it has been somehow introduced officially in the Italian system, even if not in the fully extent envisaged by Matteo Salvini.
- A thorough debate about the topic has involved in the last months tv shows, newspapers and, of course, social media.
- Aim of the work: test the *combined methodological approach* to evaluate opinion spreading about flat tax topic.

Data Collection phase

- Data are retrieved by using the current version of the free Twitter API.
- Query to search tweets has been chosen to be simply flattax . In this way, all the tweets containing that word (including #flattax) are retrieved and collected.
  Temporal window: month of May:
  1. First tweet $\rightarrowtail$ 2019-05-14 11:19:20
  2. Last tweet $\rightarrowtail$ 2019-05-27 21:05:22
- Information available are related to: text, users, replies, retweets and so on.
- Free API are not able to provide full corpus, but a sample with some restrictions.

Replies Collection phase!

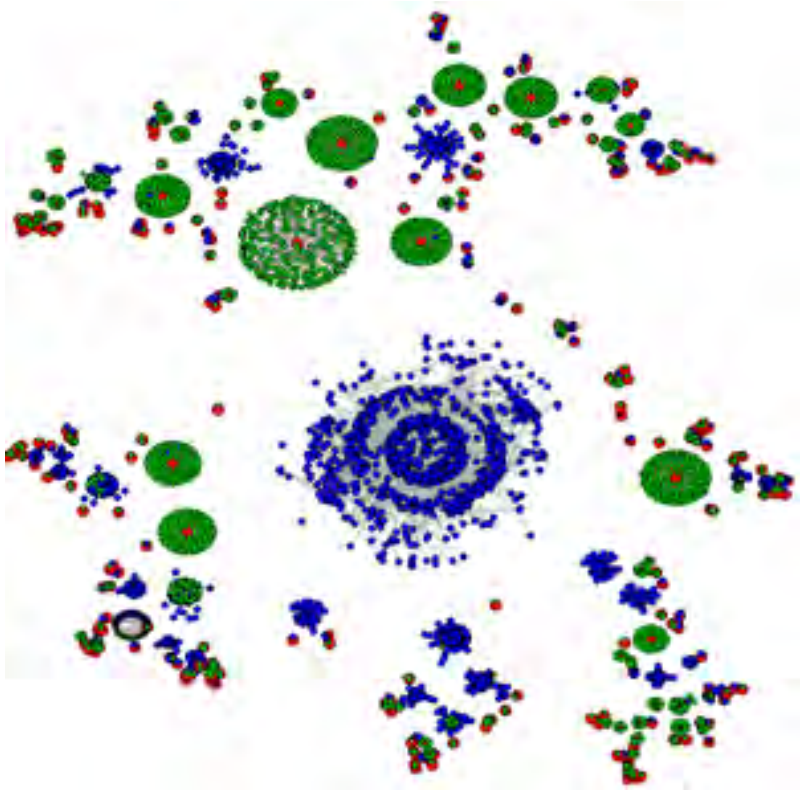
In this work a particular emphasis is given to *replies*.
The average number of replies in a random corpus is only, roughly, 1% of the total number of collected tweets.
To overcome this issue:

1. We have taken all the ID related to tweets that are reply.
2. We have taken all the username related to tweets that are replied.
3. We have done query including @username (cause each reply has to start with @username of original tweet, that is a mention).
4. Then we have filtered replies included in 3 using only IDs belonging to 1 subset of tweets.
5. Repeat procedure with new collected replies in 4 and do it several additional times iteratively, to obtain at the end a *reply chain*.

Final dataset

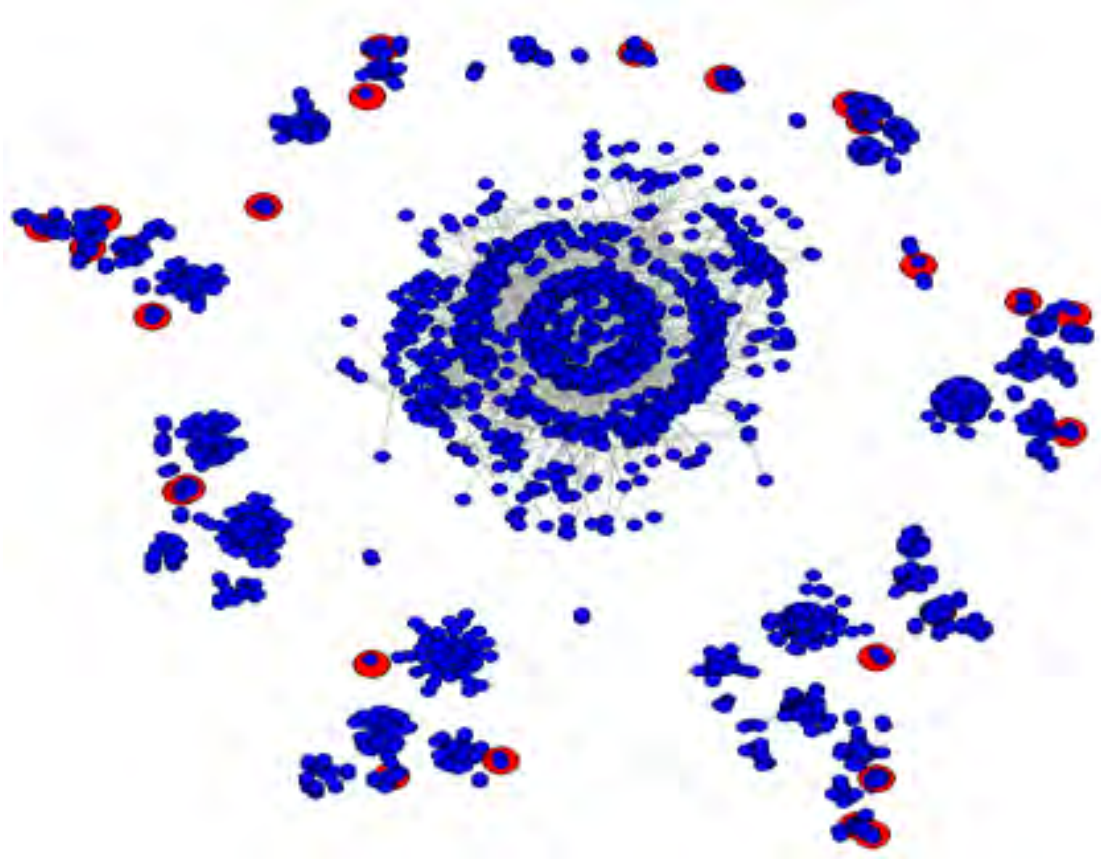
Total tweets	Original tweets	Retweets	Replies	Total N. hashtags
5994	403	2729	2862	534

Visualization of data in a graph perspective. Red: original tweets. Green: retweets. Blue: replies. Links as undirected, layout as components.



A focus on "reply" chains

Visualization of replies chains in a graph perspective. Red: original tweets. Blue: replies. Links as undirected, layout as components. Retweets and tweets with only retweets are excluded.



Finding concepts (Step 1) I

⇒ Hashtag network: Each vertex is a hashtag, undirected weighted links are co-occurrencies in original tweets (most frequent hashtags are depicted) ⇒ 12 communities of hashtags are identified

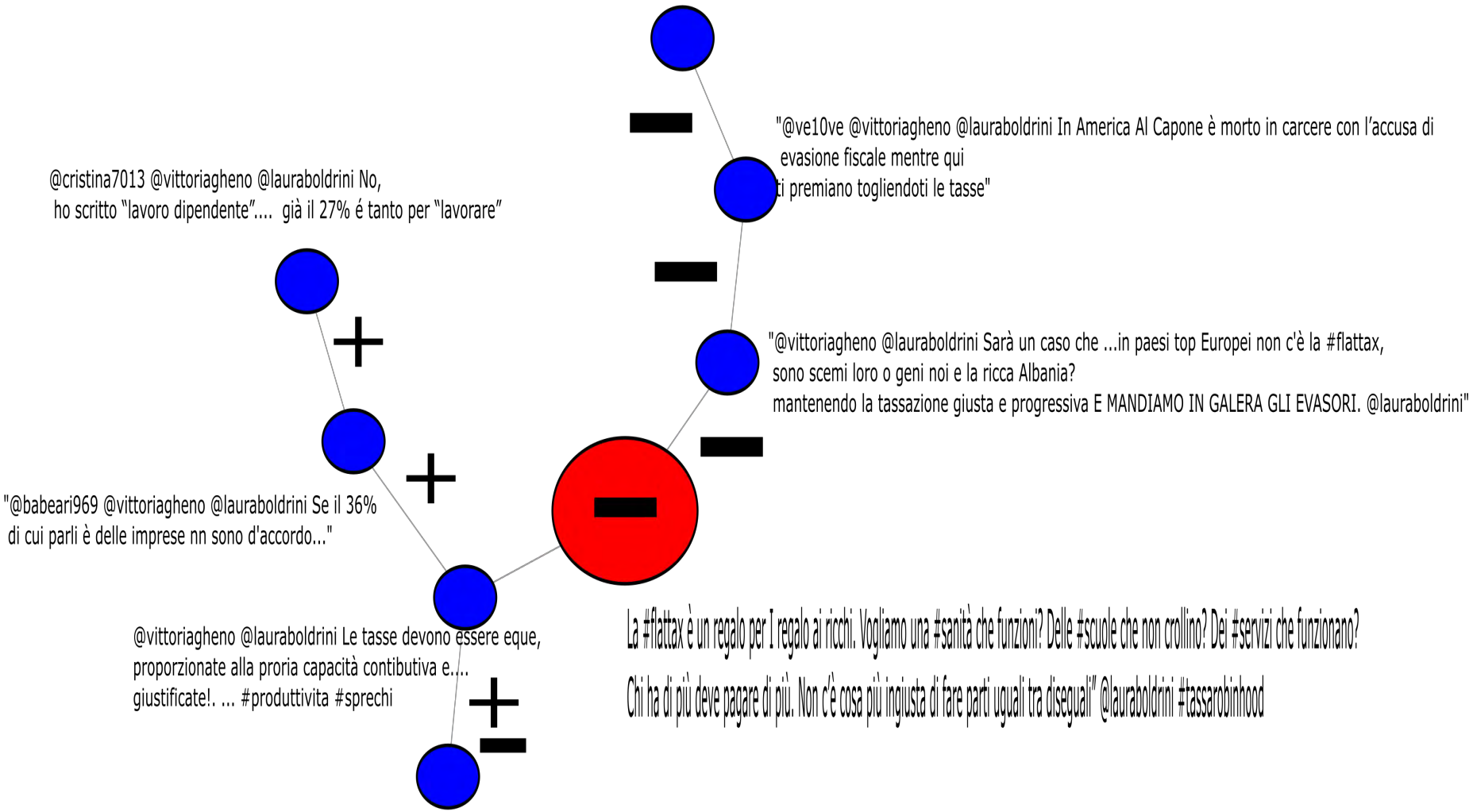
Finding concepts (Step 1) II

Concepts composition and attached sentiment (Step 2)

	Num.Hash	**positive**	**negative**	**neutral**	**Ex. hashtag**
C1	33	18%	0%	82%	#tagliamoletasse
C2	21	0%	19%	81%	#iononvotolega
C3	6	0%	0%	100%	#ansa
C4	8	0%	50%	50%	#fakenews
C5	*4*	*NA*	*NA*	*NA*	*english language*
C6	3	0%	0%	100%	#pmi
C7	3	100%	0%	0%	#votaitaliano
C8	3	66%	0%	33%	#taegdelletasse
C9	2	100%	0%	0%	#stoconsalvini
C10	2	0%	0%	100%	#fisco
C11	*7*	*NA*	*NA*	*NA*	*french language*
C12	*4*	*NA*	*NA*	*NA*	*english language*

Conditional sentiment analysis (step 3)

A case of reply signed chain



Comparing signed networks (step 4)

Summary of the n signed networks within each concept $\rightarrow$ spreading behaviour

	Sign	**Retweets**	**Reply +**	**Reply -**
Concept1	+	1246	40	126
Concept2	-	816	25	41
Concept3	Neutral	42	15	82
Concept4	-	130	20	9
Concept6	+	20	0	0
Concept7	+	6	1	1
Concept8	+	1	0	0
Concept9	+	1	0	0
Concept10	Neutral	0	0	0
	Total:	**2262**	**101**	**259**

Concluding remarks

- A first attempt to combine SNA and SA to analyze structure of opinion spreading on Twitter.
- the approach leads to a signed networks describing the structure of retweet and reply interactions of polarized concepts related to a trending topic.

Open issues

- Time-consuming Sentiment analysis: as long as the reply chain is extended we have to run several times a sentiment analysis algorithm.
- How "automatic" should sentiment analysis be?
- Message-based approach: e.g., the level of "influenceness" of the original user producing the tweet is not considered
- Several approaches to analyze the obtained signed networks
- Human judges, expert of the topic, should be used to estimate precision and recall.

Future improvements

- Combining user-based analysis in the Sentiment Analysis step
- Using ERGM on signed networks to model the way each concept spreading structure