Advanced Data Mining and Machine Learning
Text Mining Case Studies

Academic Year 2022-2023

Alessandro Renda

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

UNIVERSITÀ DI PISA

Outline

- An introduction to social sensing (slides from professor P. Ducange)
- A case study: stance towards vaccination in Italy
- The stance classification platform
- User-oriented and geospatial analysis
- Concept drift analysis in tweets classification
- A case study: stance towards green-pass (EU digital COVID certificate) in Italy

Outline

- An introduction to social sensing (slides from professor P. Ducange)
- A case study: stance towards vaccination in Italy
- The stance classification platform
- User-oriented and geospatial analysis
- Concept drift analysis in tweets classification
- A case study: stance towards green-pass (EU digital COVID certificate) in Italy

Social Media and Social Networks

- ***Social media*** is content that users upload, namely a blog, a video, a slideshow, and so on
- ***Social media*** is also the «venue» where content are published and shared
- ***Social networking*** is the practice of connecting and engaging with others on the social media sites.

Image extracted from "Social Media Landscape 2017" available at: https://fredcavazza.net/2017/04/19/social-media-landscape-2017/

Twitter as a micro blogging platform

- ***Why*** people actively use Twitter:
  - to report (personal or public) real-life events
  - to express their opinion on a given topic, through a ***public message***
  - **…**

- ***How*** people actively use Twitter:
  - The technical term for a “tweet” or “post” is ***Status Update Messages*** (SUMs)
    - SUMs may include ***meta-information*** such as timestamp, geographic coordinates (latitude, longitude), username, links to other resources, hashtags, and mentions
    - Several SUMs referring to a certain topic or related to a limited geographic area may provide, if properly analyzed, great deal of ***valuable information*** about an ***event*** or a ***topic.***

Sensor Networks

- A Sensor Network consist of a set of physical sensing devices
  - either inside the phenomenon or close to it
- Possibly wired or wireless sensors
- Architectures and algorithms for handling and leveraging information are widely investigated

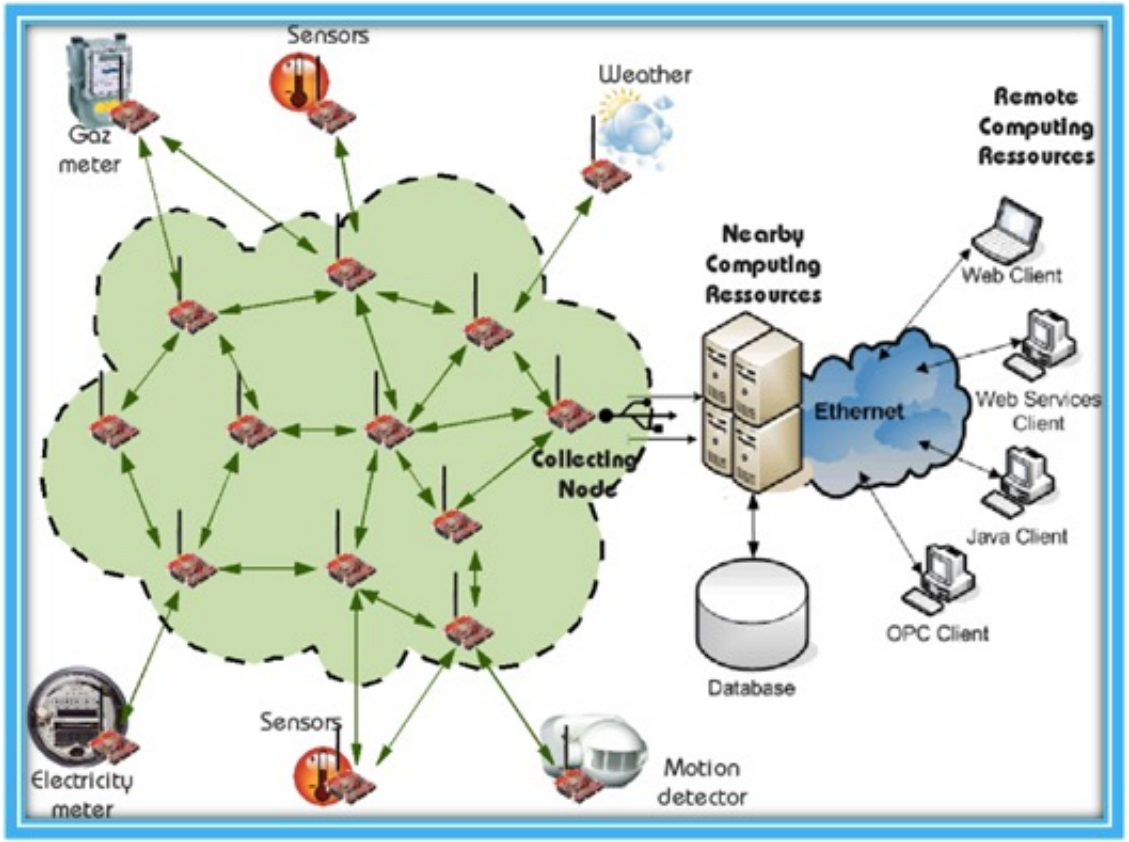
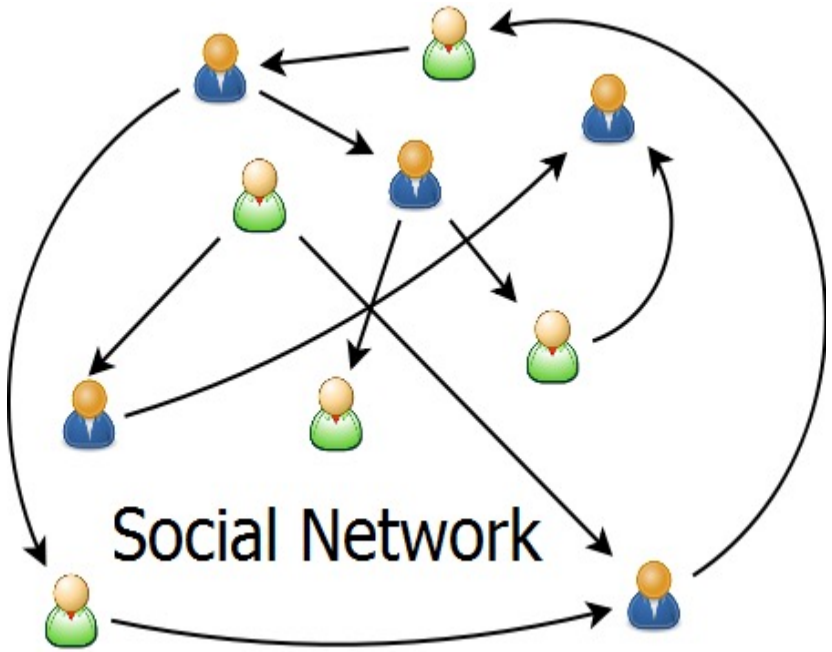


Image extracted from: http://omnet-tutorial.com/omnet-code-for-wireless-sensor-networks/

Social Sensing

- We may regard social network users as ***social or human sensors***
- Status Update Messages (SUMs) represent ***sensor information***
- Social sensors can give indications regarding ***population***
- Social sensors and “classical sensors” can be ***integrated***



Social Sensing Applications

- Real-time traffic and urban activity monitoring
- Prediction and analysis of events
- Feedbacks about the satisfaction of a service
- Discover trends in public emotion
- Analysis and Prediction of Fashion Trends
- Prediction of Election Results

Sentiment Analysis, Opinion Mining, Stance Detection

Sentiment Analysis
Focus on the sentiment.
Determine whether a piece of text is positive, negative, neutral, or the specific emotion or the polarity

Opinion Mining
Focus on the opinion.
Determine speaker's opinion along with the relative target.

Stance Detection
Target is pre-chosen.
Determine the opinion towards it, no matter if it is explicitly mentioned in the text, no matter which is the underlying sentiment

Outline

- An introduction to social sensing
- A case study: stance towards vaccination in Italy
- The stance classification platform
- User-oriented and geospatial analysis
- Concept drift analysis in tweets classification
- A case study: stance towards green-pass (EU digital COVID certificate) in Italy

Case study: stance towards vaccination

- Online **debates and discussion groups** over social networks about vaccination topic
  e.g. alleged connection between **autism** and **MMR vaccine**
- News and opinions spread in social networks **influence individual behaviors and sentiments**

- March 2017: Italian Ministry of Health detected an overall **drop in vaccination coverage**
- Risk of **re-emergence of eradicated diseases** or a **raise in the incidence rate**

Idea presented during a congress in Erice (TP) in **August 2018**

Case study: stance towards vaccination

- Automatic monitoring of the **stance of Twitter users** towards vaccination topic

- Design and implementation of a **text-mining system** for stance classification task: determine if a tweet is
  - *In favor* of vaccination
  - *Not in favor* of vaccination
  - *Neutral*

- **Detection of opinion shifts**

- Public Healthcare Organizations could take the pulse of public opinion and promote countermeasures

News of the world

AJPH
A PUBLICATION OF THE AMERICAN PUBLIC HEALTH ASSOCIATION

Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate

David A. Broniatowski PhD, **Amelia M. Jamison** MAA, MPH, **SiHua Qi** SM, **Lulwah AlKulaib** SM, **Tao Chen** PhD, **Adrian Benton** MS, **Sandra C. Quinn** PhD, and **Mark Dredze** PhD (show fewer authors)

[+] Author affiliations, information, and correspondence details

Accepted: May 22, 2018 Published Online: August 23, 2018

David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze, 2018: Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate American Journal of Public Health 108, 1378_1384, https://doi.org/10.2105/AJPH.2018.304567

Text Mining Perspective

- **Extraction of meaningful information** out of tweets, unstructured natural language text

- Challenging setting:
  - Typically, **unstructured and irregular text,** with **limited length**
  - **Informal words**, colloquial, idiomatic expressions, misspellings or grammatical errors
  - Acronyms, hashtags, URLs and mentions
  - **Ambiguity** of opinion, sarcasm, irony

Noisy source of information

Outline

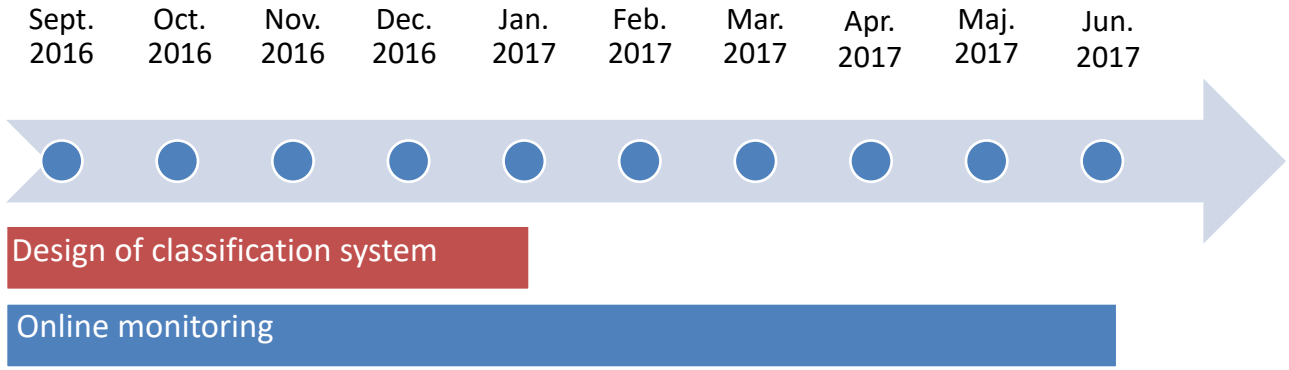
- An introduction to social sensing (slides from professor P. Ducange)
- A case study: stance towards vaccination in Italy
- The stance classification platform
- User-oriented and geospatial analysis
- Concept drift analysis in tweets classification
- A case study: stance towards green-pass (EU digital COVID certificate) in Italy

Gathering Twitter Data

- Focus on **Italian setting**
- Use of 37 **vaccine-related keywords / hashtags**
  - *... vaccino – big pharma – autismo – morbillo – #iovaccino – #libertadiscelta - ...*
- Querying Twitter API and external libraries

- Dataset of **112.937 tweets**
- Time interval of 10 months: **September 2016 - June 2017**

Training stage and monitoring campaign



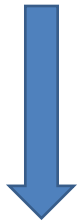
- Design of a proper **automatic classification system**
  - Training dataset of 693 manually labelled tweets
  - Experimental campaign to find the best text representation and classification model: 10-fold stratified cross-validation for models comparison
- Online monitoring
  - Its use for uncovering **trends of public opinion** over time

Building the model: Comparing different approaches

Table 3
Average results obtained by using the different approaches discussed in the text.

Classifier	Class	F-measure	Precision	Recall	AUC	Accuracy
BOW + SVM_All	*Not in favor*	0.60	62.6%	56.6%	0.73	65.4%
	In favor	0.65	64.5%	65.5%	0.74	
	Neutral	0.71	68.6%	74.0%	0.80	
BOW + SVM_2000	*Not in favor*	0.59	61.5%	56.2%	0.73	64.8%
	In favor	0.64	63.2%	63.9%	0.74	
	Neutral	0.72	69.4%	74.4%	0.81	
BOW + FASTTEXT + SVM	*Not in favor*	0.59	57.9%	60.3%	0.75	64.2%
	In favor	0.73	73.3%	72.6%	0.82	
	Neutral	0.61	62.1%	60.4%	0.72	
BOW + GLOVE + SVM	*Not in favor*	0.56	59.5%	53.0%	0.74	62.2%
	In favor	0.70	66.9%	72.1%	0.79	
	Neutral	0.61	59.9%	61.6%	0.71	
BOW + W2V + SVM	*Not in favor*	0.59	61.1%	56.6%	0.73	63.7%
	In favor	0.72	68.9%	74.9%	0.81	
	Neutral	0.60	60.7%	60.0%	0.72	
FASTEXT + CNN	*Not in favor*	0.57	57.8%	57.9%	0.69	62.9%
	In favor	0.63	64.3%	62.7%	0.70	
	Neutral	0.68	69.6%	68.0%	0.77	
GLOVE + CNN	*Not in favor*	0.55	54.8%	56.6%	0.67	60.5%
	In favor	0.63	64.5%	62.4%	0.71	
	Neutral	0.63	65.1%	62.2%	0.73	
W2V + CNN	*Not in favor*	0.57	57.2%	58.0%	0.69	62.5%
	In favor	0.62	62.9%	61.6%	0.70	
	Neutral	0.69	70.4%	68.1%	0.77	

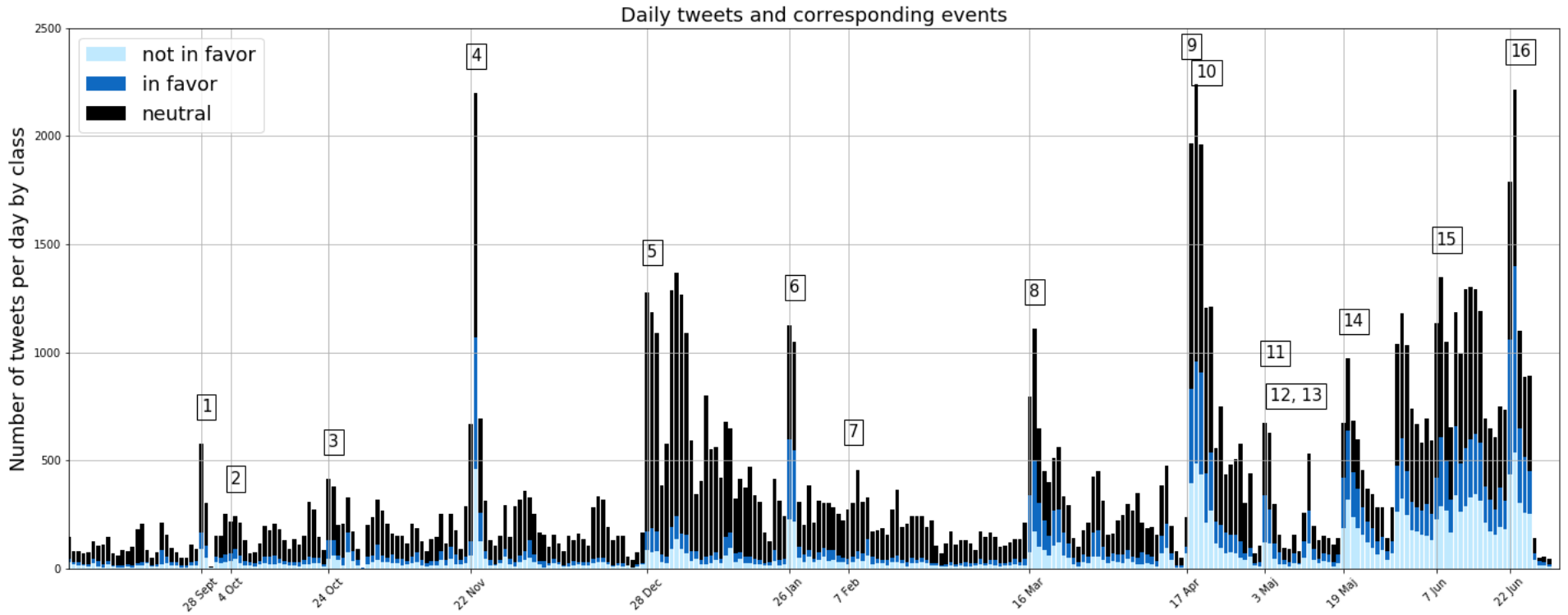
- **BOW+SVM_All** as control model, pairwise **Wilcoxon signed-rank test**
- **BOW+SVM_2000** statistically equivalent



- Selected as the most suitable scheme

Online Monitoring

- Usage of the classification system on the unlabelled tweets.
  - 112.244 tweets, time span of 10 months

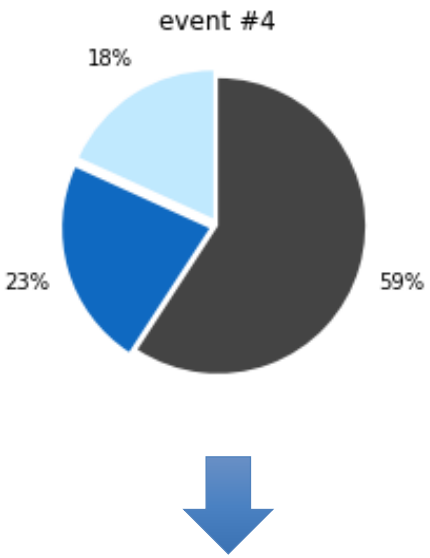


Occurrence of socio-political vaccine-related events

1. Cancellation of the projection of the documentary film “Vaxxed: from cover-up to catastrophe” in the Italian Republic Senate on September 28th, 2016
2. Expected projection of the documentary film “Vaxxed: from cover-up to catastrophe” in the Italian Republic Senate on October 4th, 2016
3. Speech by President of Italian Republic about vaccines on October 24th, 201
4. Approval of the law establishing vaccination requirements for school children in Emilia Romagna Region, Italy, approved on November 22nd, 2016
5. Death of a schoolteacher for meningitis in Rome, Italy, news of December 28th, 2016
6. Agreement between Italian Health Minister and Italian Regions about vaccinations requirement on January 26th, 2017
7. Cancellation of the projection of the documentary film “Vaxxed: from cover-up to catastrophe” at the European Parliament on February 7th, 2017
8. Increase of 230% cases of measles in Italy, news of March 16th, 2017
9. Italian TV show Report focusing on vaccines cause controversy on April 17th, 201
10. Fake vaccinations in the Italian city of Treviso, news of April 19th, 2017
11. Fake vaccinations in the Friuli Region, Italy, news of May 3rd, 2017
12. NY Times against Italian political party against vaccines, news of May 4th, 2017
13. 5 times increase in measles cases in Italy in April 2017, news of May 4th, 2017
14. Approval of the decree on vaccinations requirement (12 vaccines) in Italian kindergartens on May 19th, 2017
15. President of Italian Republic signs the decree about 12 vaccinations requirement in Italian schools on June 7th, 2017
16. Kid sick of leukemia died for measles in Monza, Italy, news of June 22nd, 2017

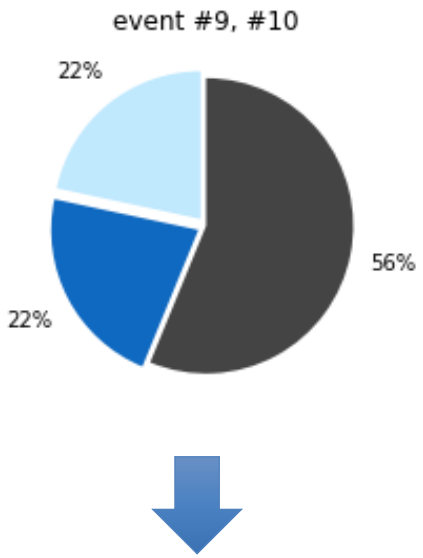
Occurrence of socio-political vaccine-related events

4. Approval of the law establishing vaccination requirements for school children in Emilia Romagna Region, Italy, approved on November 22nd, 2016

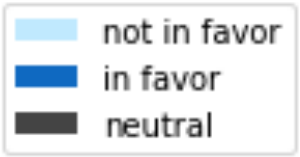


Slightly biased towards *in favor of vaccination* (+5%).

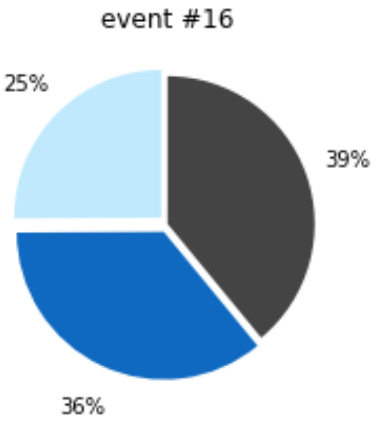
9. Italian TV show Report focusing on vaccines cause controversy on April 17th, 201
10. Fake vaccinations in the Italian city of Treviso, news of April 19th, 2017



Overall *neutral*



16. Kid sick of leukemia died for measles in Monza, Italy, news of June 22nd, 2017

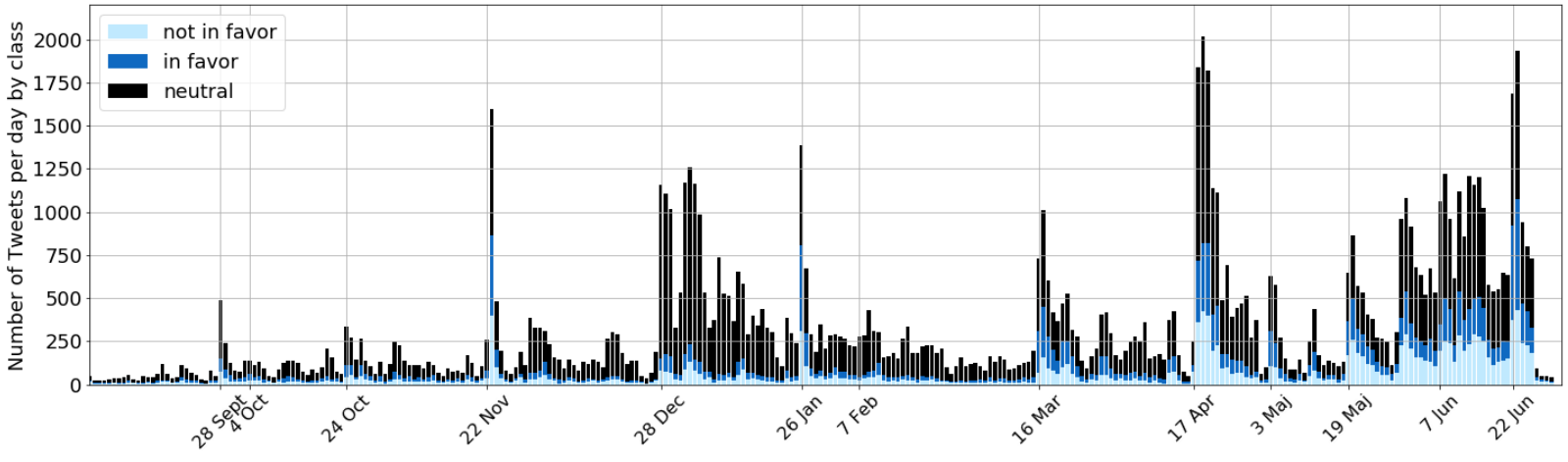


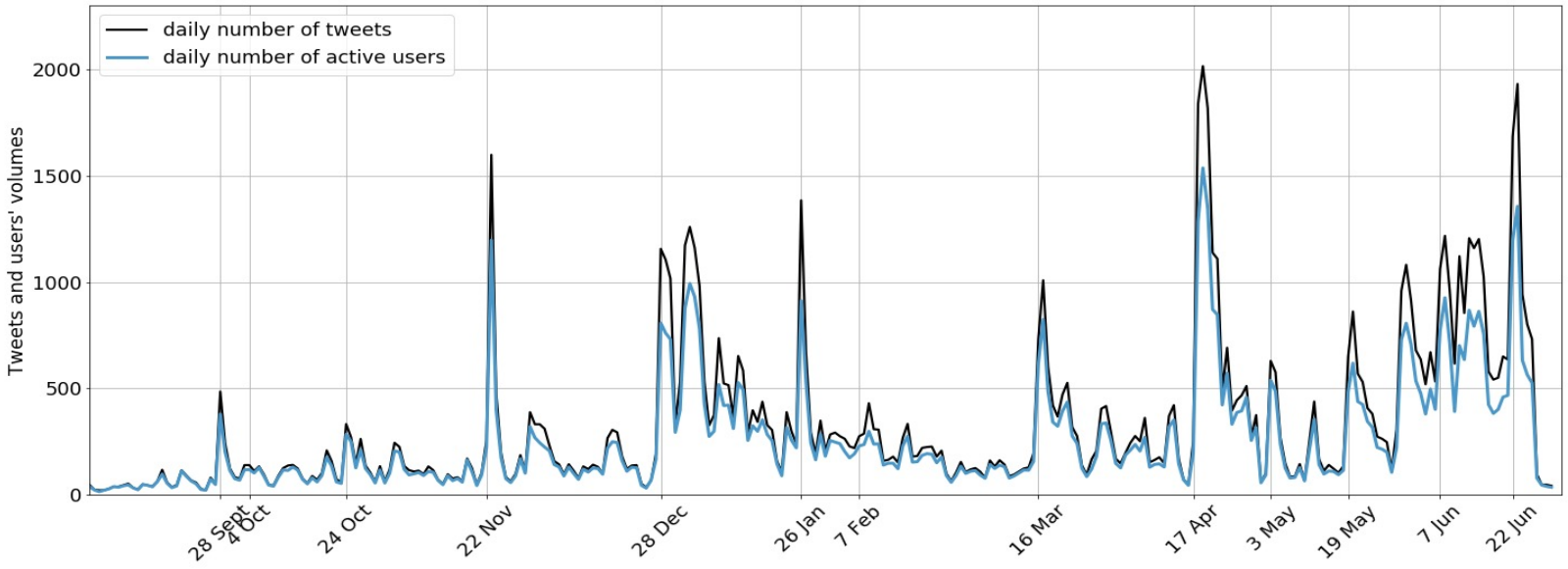
Considerably biased towards *in favor of vaccination* (+11%)

Outline

- An introduction to social sensing
- A case study: stance towards vaccination in Italy
- The stance classification platform
- User-oriented and geospatial analysis
- Concept drift analysis in tweets classification
- A case study: stance towards green-pass (EU digital COVID certificate) in Italy

Stance Detection System

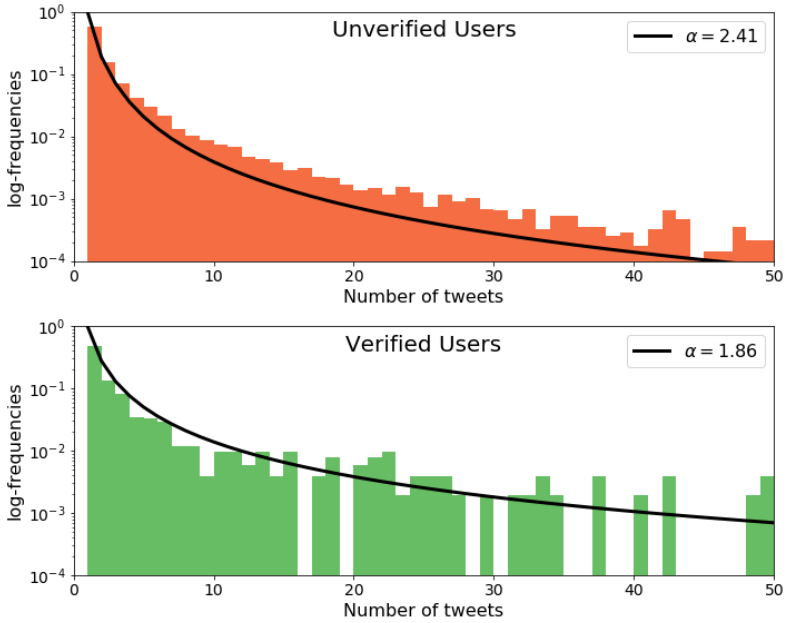




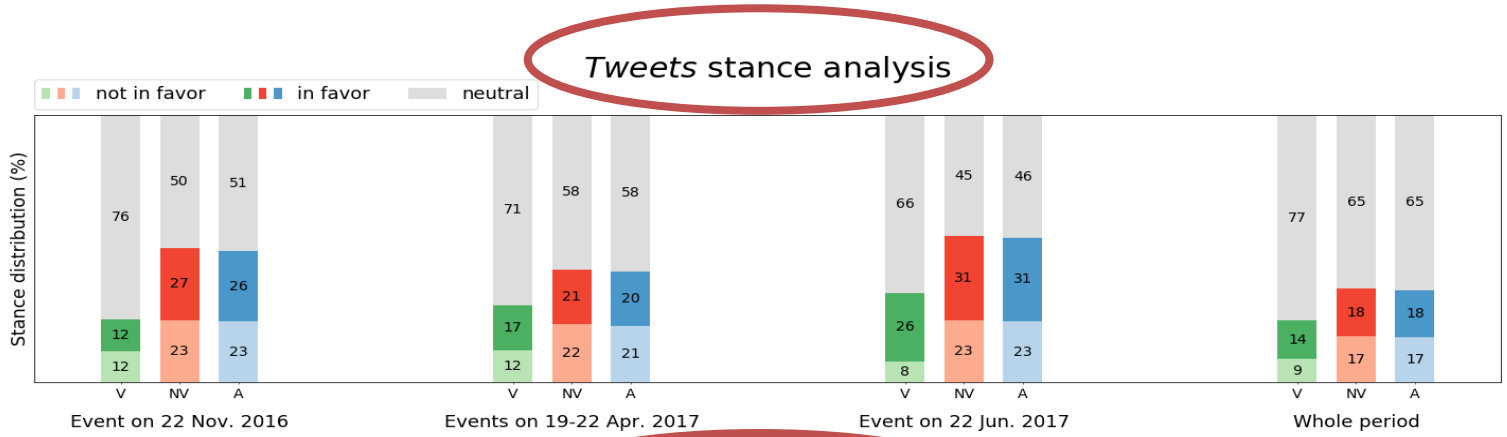
Observation

- Is stance analysis on tweets representative of public opinion?

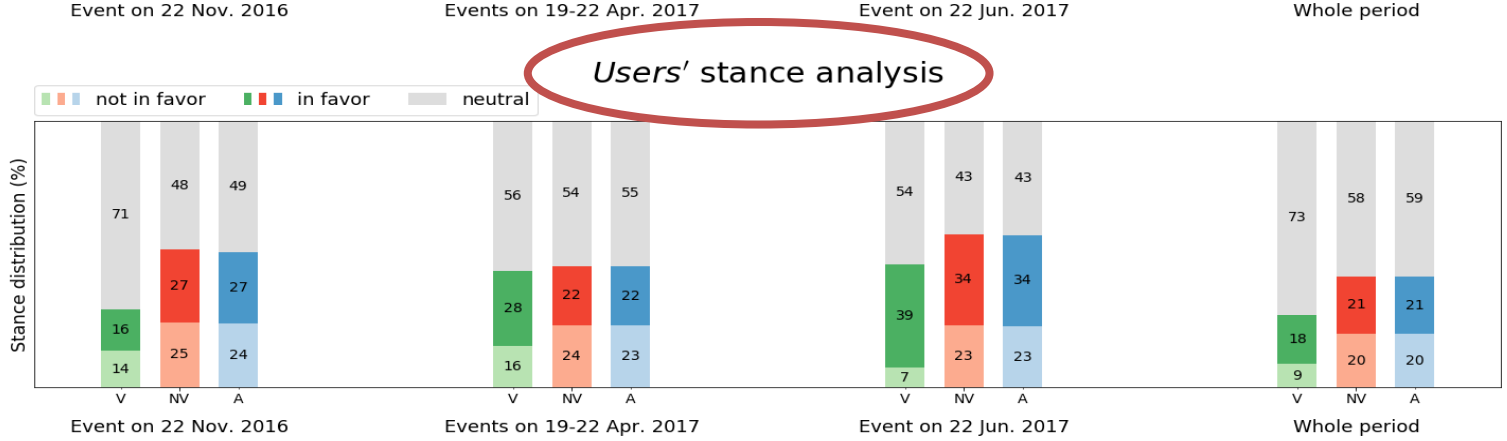
User's Stance Detection



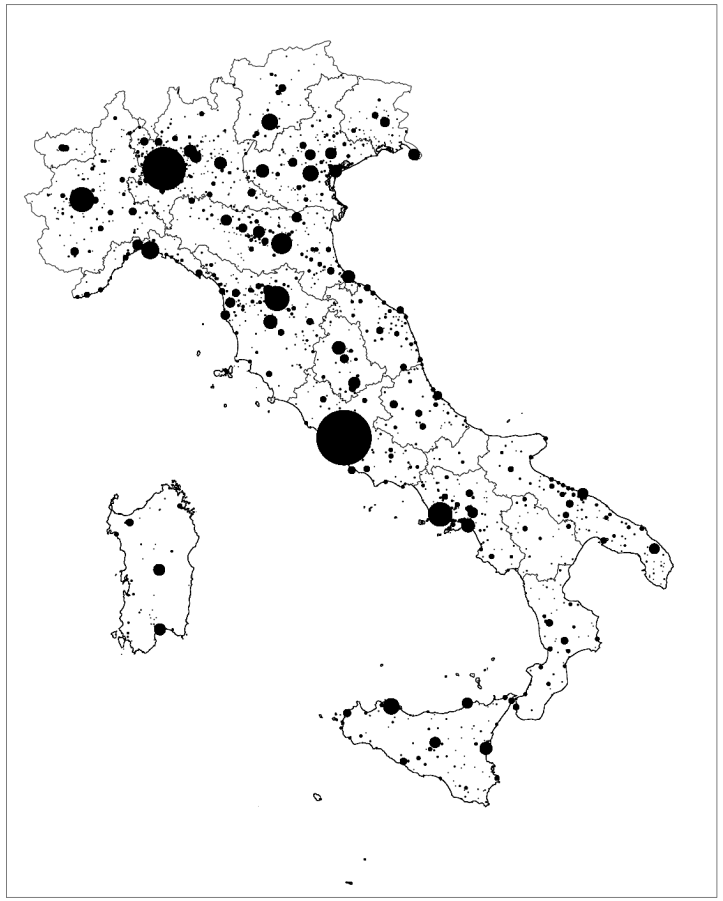
- Different activity behavior between:
  - **verified** (mainly news-wires and news agencies)
  - **unverified** groups (mainly individuals)

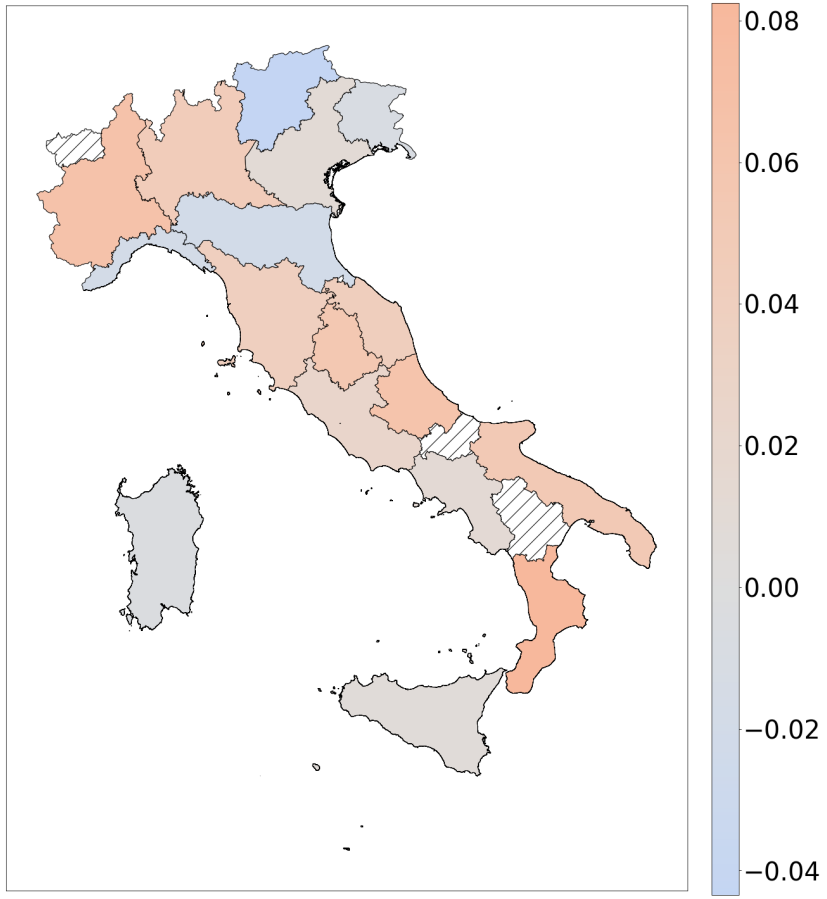


Sharper picture of public opinion



User's Stance Detection: Geospatial Analysis





Exploitation of user location info:

Regional-scale analysis of activity and stance

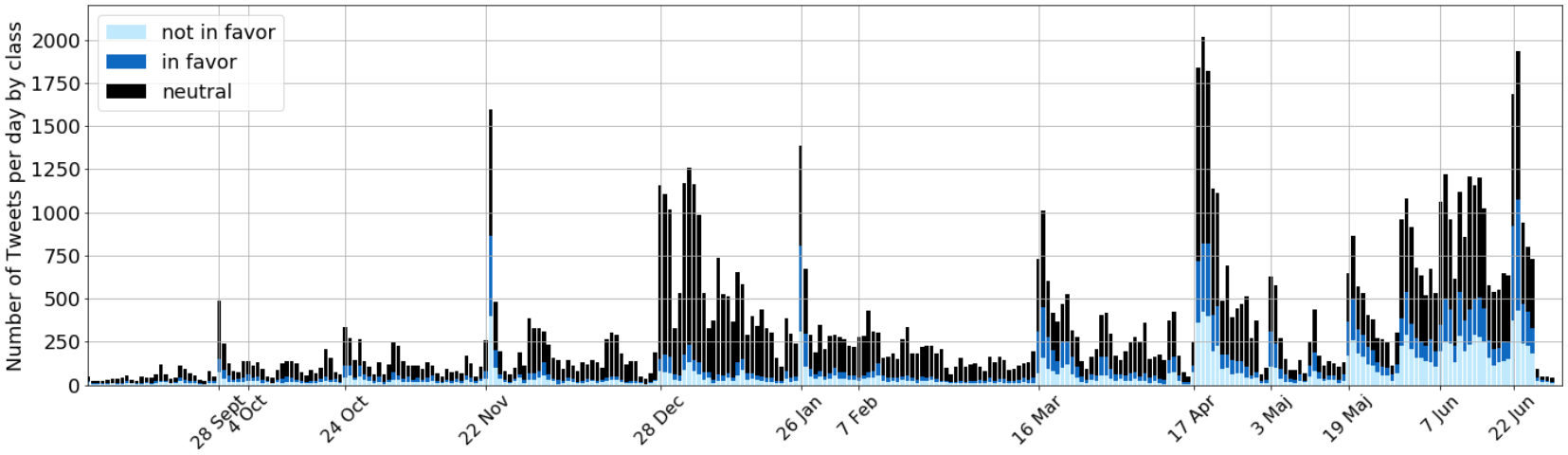
$$\text{RegionalScore}_r = \frac{\text{InFavor}_r - \text{NotInFavor}_r}{\text{Total}_r}$$

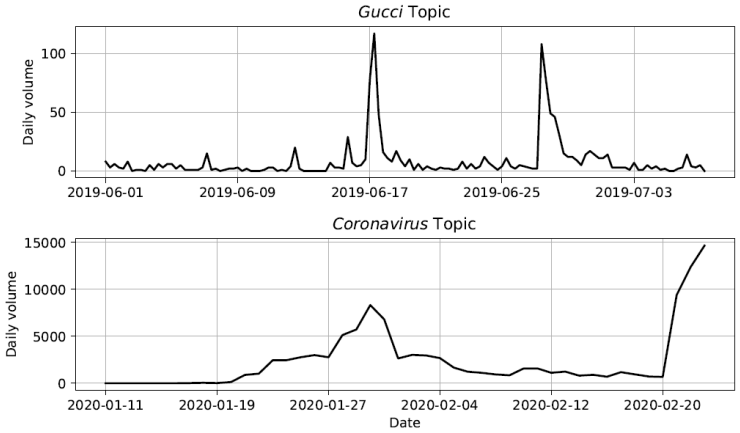
Outline

- An introduction to social sensing
- A case study: stance towards vaccination in Italy
- The stance classification platform
- User-oriented and geospatial analysis
- Concept drift analysis in tweets classification
- A case study: stance towards green-pass (EU digital COVID certificate) in Italy

Addressing Event-Driven Concept Drift

- **Observation**
  - Twitter stream may be affected by concept-drift driven by *real-world* events

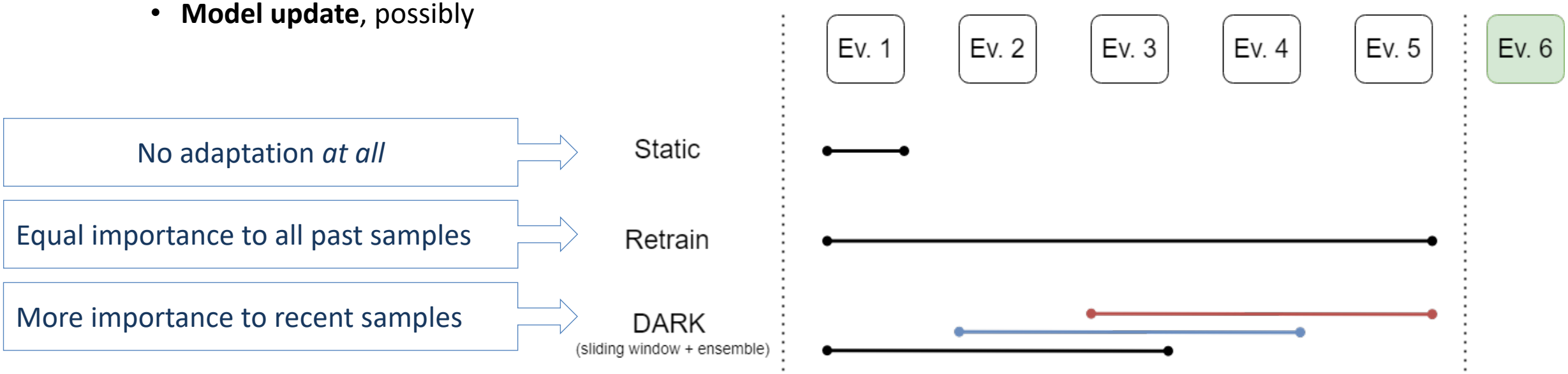




- **Research questions**
  - How to deal with concept drift in a tweet classification task?
  - Is an *initial* classification model suitable for long-term online monitoring?

Stance Detection from Twitter Stream

- **Experimental Setup**
  - Empirical comparison of different **learning schemes**
  - Extended monitoring campaign (~3 years)
  - At each event we label some tweets. Data available for:
    - **Performance evaluation** over time
    - **Model update**, possibly

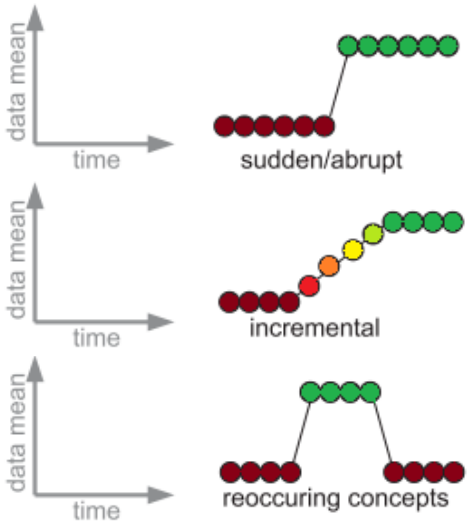


Stance Detection from Twitter Stream

- **Observation**:
  - **Concept drift** definition: given two timestamps $t0, t1$, it holds:

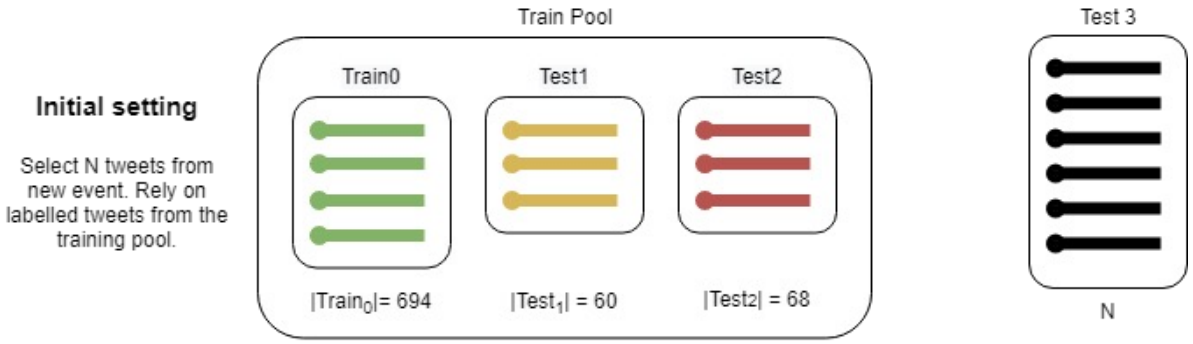
$$\exists \boldsymbol{X}: \; p_{t0}(\boldsymbol{X}, y) \neq p_{t1}(\boldsymbol{X}, y)$$

    where p_{ti} is the joint probability distribution between input and output
    Concept drift may appear in **different forms**
  - In the **dynamic Twitter setting**:
    - Recent data are not necessarily the most relevant ones
    - Old data are not necessarily to be discarded



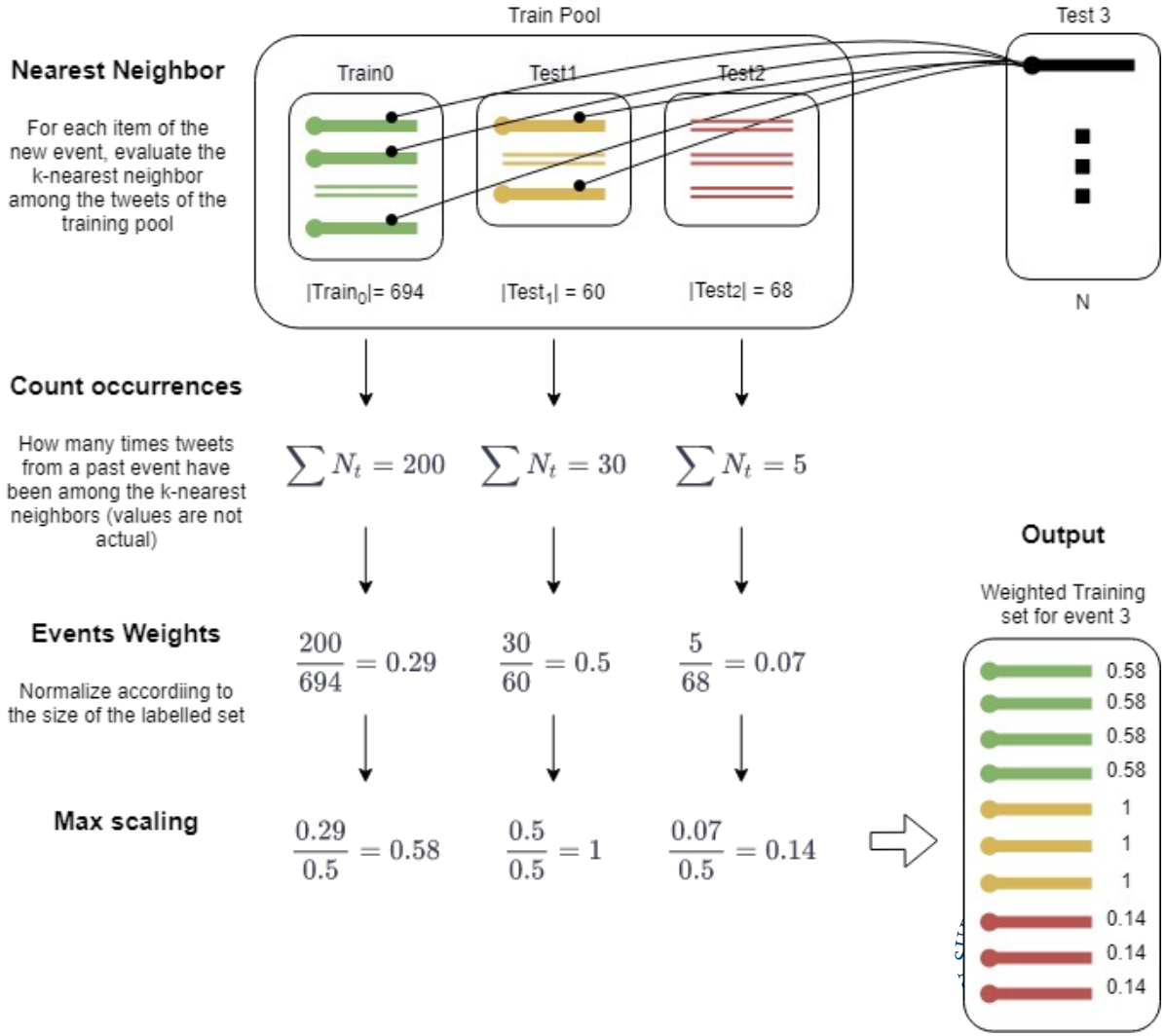
- **Proposed approach**: **Semantic-Aware Learning scheme**
  - Similar to the *Retrain* learning scheme
  - Weigh each sample from past events according to the *semantic similarity* between the newly collected event and each past event

Stance Detection from Twitter Stream



- Similarity evaluation based on numerical representation of tweets obtained with AlBerto
  - Adaptation of **BERT** language model for Italian
  - Tailored for the Twitter environment

*How many tweets from a past event are among the **k-NN** of each of the newly collected tweets?*



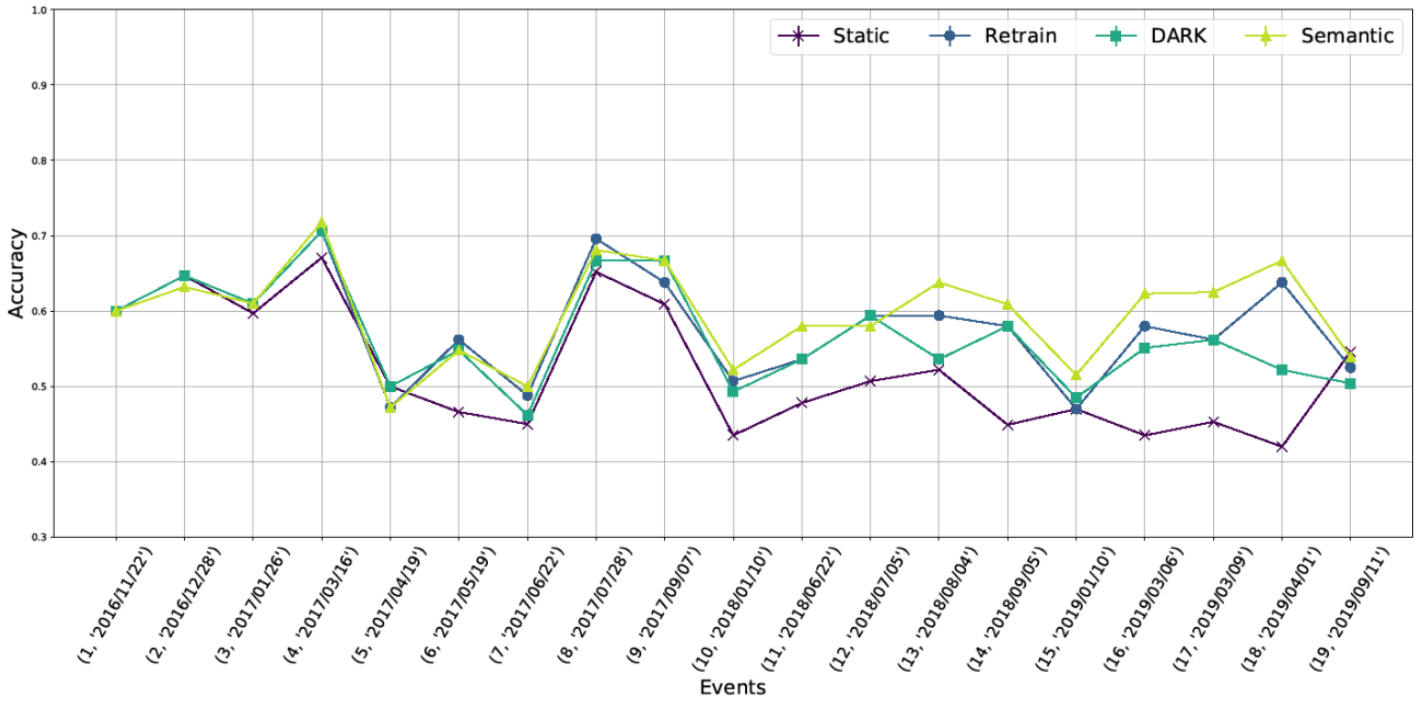
Stance Detection from Twitter Stream

- **Qualitative evaluation**: examples of neighboring tweets

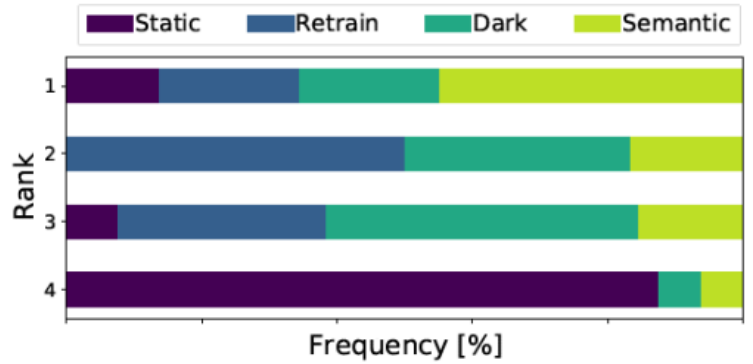
	Aumento (preoccupante) dei casi di morbillo in Italia. No, ma l'importante è che le merendine non contengano olio di palma
NN-1	+230% di casi di #morbillo in Italia in un solo anno. Ma mi raccomando, continuate a dire che i vaccini fanno male
NN-2	Preoccupante boom dei casi di #morbillo in #Italia a fronte del calo dei #vaccini. Chi fa propaganda anti-vaccinazione è un delinquente.
NN-3	@RaiRadio2 come dimostrereste che l'aumento dei casi di morbillo nel 2017 "é sicuramente dovuto" alla diminuzione delle vaccinazioni?
NN-4	Triplicati i casi di #morbillo in Italia, colpa del crollo dei #vaccini http://agi.it/salute/2017/03/17/news/triplicati_- i_casi_di_morbillo_in_italia1592660/
NN-5	Non vaccinate i vostri figli, mi raccomando! Intanto in Italia c'è un grande aumento dei casi di morbillo! Bravi!

	a scuola a Pagani
	tti i bambini potranno andare a scuola anche senza
	ettembre
	cuola - Focus vaccini
NN-4	VACCINI. DE LUCA, 'IN CAMPANIA BIMBI NON VACCINATI NON AMMESSI A SCUOLA' | Tv7 Benevento
NN-5	Per fare andare a scuola i bambini non vaccinati non serve un decreto legge, basta vaccinarli...

Stance Detection from Twitter Stream



Rank frequencies of accuracy values scored:



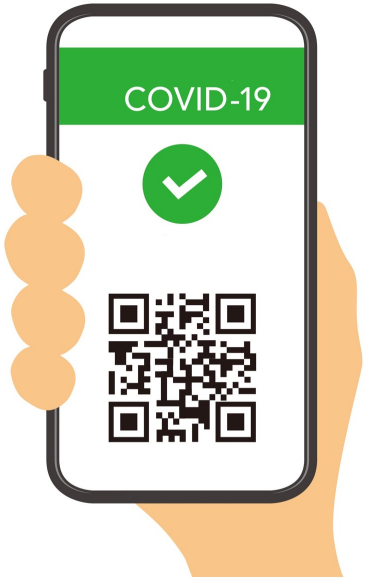
- High variability of accuracy scores between events
- *Static* approach is not suitable for long-term monitoring campaign
- *DARK* and *Retrain* achieve comparable performance
- Slightly better results by *Semantic* approach

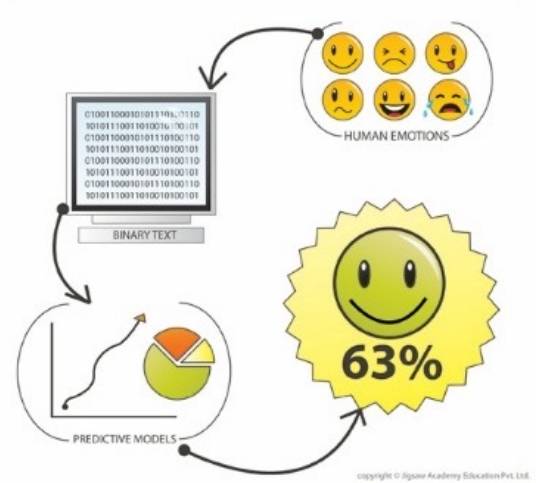
Outline

- An introduction to social sensing
- A case study: stance towards vaccination in Italy
- The stance classification platform
- User-oriented and geospatial analysis
- Concept drift analysis in tweets classification
- A case study: stance towards green-pass (EU digital COVID certificate) in Italy

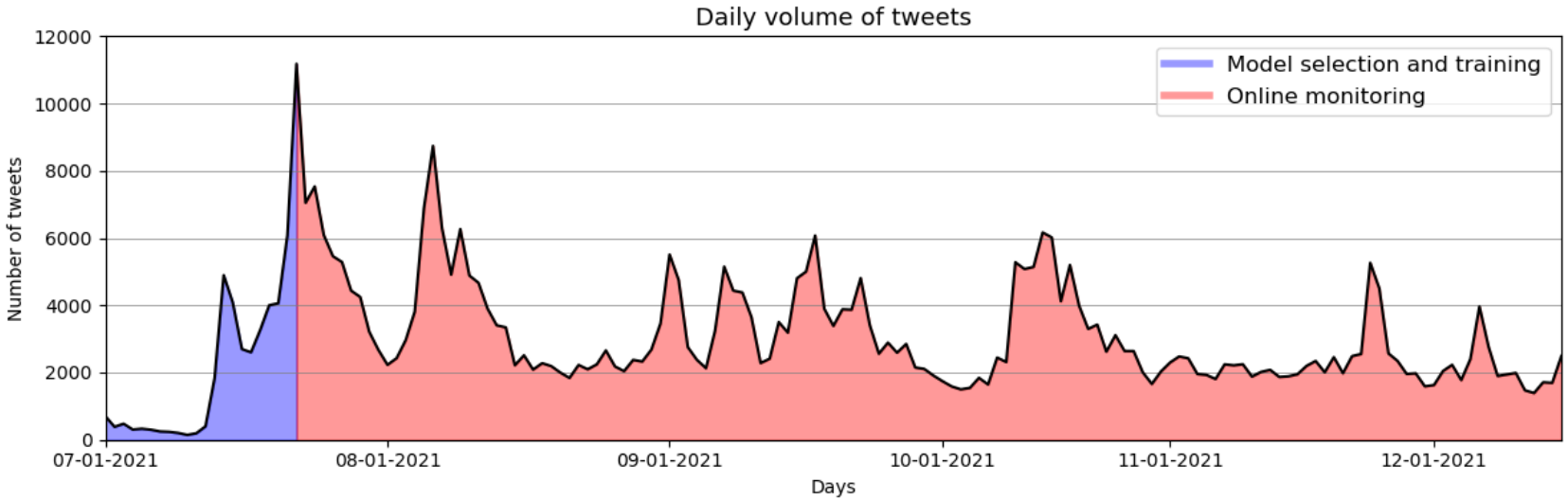
The case of Green Pass in Italy

- **Objectives**:
  - Automatically monitoring the public opinion (stance) on ***social media*** regarding Green Pass in Italy
  - ***Nowcast*** public opinion providing a fast, real-time, low-cost, and easy alternative to traditional polls and surveys
  - Analyze the occurrence of concept drift and define some strategies for managing it.





The Dataset



- **Training set time window**
  01/07/2021 - 22/07/2021
- Training set size:
  - **2400 tweets**
  - Balancing among classes

- **Fetching stage:**
  - Keywords: “greenpass” or “green pass”
  - Italian Language
  - Italian Region

around **500K tweets**

- **Pre-processing stage:**
  - Keep only Italian tweets
  - Removing URL, mentions, emoticons
  - Replacing multiple spaces with single space
  - Removing punctuation marks
  - Lower case

Model Selection

- **Traditional text elaboration pipeline,** with Bag-of-Words representation and TF-IDF scheme
- Results of the 10-fold stratified cross-validation
- ComplementNB as the most suitable model, as it achieves good performances and supports ***online learning***

Classifier	Class	Precision	Recall	F_1-score	F_1^{PN}	Accuracy (%)
ComplementNB	Neutral	0.6447	0.5328	0.5834	0.6745	64.76
	Positive	0.6035	0.7850	0.6824		
	Negative	0.7192	0.6212	0.6667		
LogisticRegression	Neutral	0.5928	0.5700	0.5812	0.6676	63.94
	Positive	0.6494	0.6800	0.6643		
	Negative	0.6742	0.6675	0.6709		
SVM	Neutral	0.5835	0.6059	0.5945	0.6626	63.94
	Positive	0.6542	0.6534	0.6538		
	Negative	0.6844	0.6587	0.6713		

Concept Drift Analysis

- We assume that a collection of ***tweets*** can be also ***manually labelled*** by a user throughout the ***monitoring*** campaign.
- Since ***labeling*** each single tweet is ***impractical***, we resort to a ***strategy*** for labeling a representative set of tweets.
- We adopt a ***buffer*** of 200 elements:
  - In a given time period, some tweets are randomly selected
  - The user can label the tweets by confirming or correcting the predicted label
  - Whenever the buffer is full and re-labeled, the chunk of tweets is first used for testing the classification system performance and then, possibly, to update it.

1600 *additional* tweets labelled between 07/2021 and 12/2021

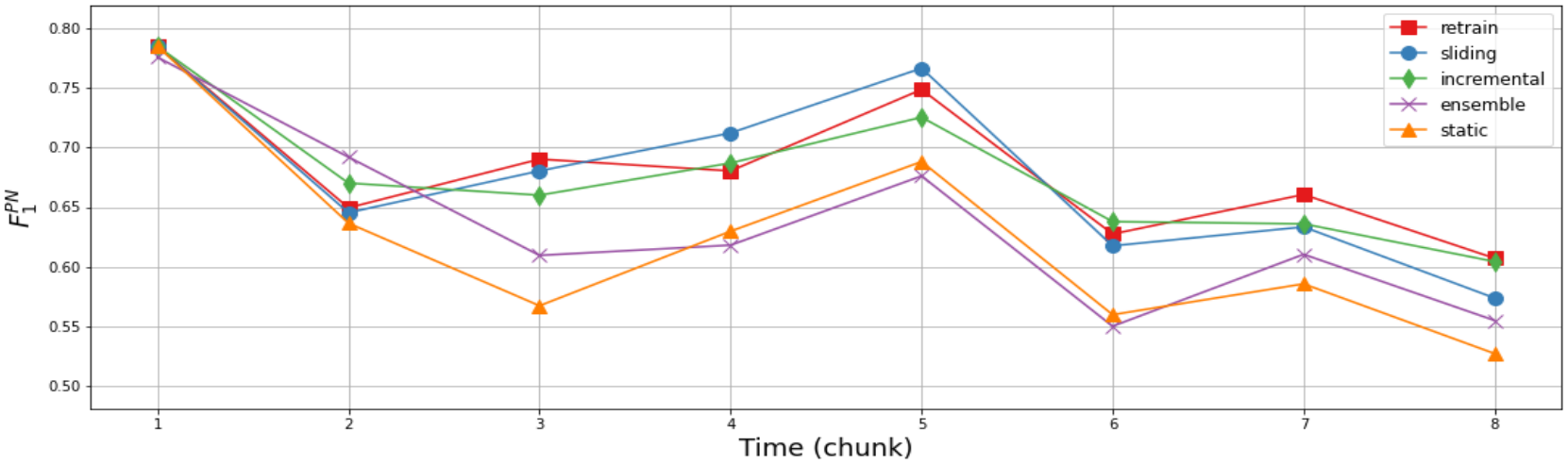
CLASS DISTRIBUTION OF CHUNKS OF TWEETS AFTER MANUAL LABELLING.

	Positive	Negative	Neutral
Chunk 1	101	79	20
Chunk 2	41	56	103
Chunk 3	41	64	95
Chunk 4	81	63	56
Chunk 5	75	68	57
Chunk 6	57	66	77
Chunk 7	65	69	66
Chunk 8	72	74	54

Concept Drift Analysis

- ***static*** (baseline): the classification system trained on the original training set
- ***retrain***: the training set is extended with each new labelled chunk of tweets and the whole classification pipeline is retrained from scratch
- ***sliding***: analogous to retrain, but *oldest* tweets are removed from the training set
- ***incremental***: the classification model is updated based on the new labelled chunk of tweets with a *partial fitting*
  - unlike the *retrain* approach, the old model is not replaced with a new one trained from scratch; rather its parameters, namely its internal statistics, are updated considering the new data distribution.
  - only the classifier is updated, whereas the attribute space, namely the vocabulary generated based on the initial TF-IDF vectorization, is left unchanged throughout the online monitoring
- ***ensemble***: the prediction of the three *static* classifiers (SVM, Logistic Regression, ComplementNB) are combined based on a majority voting policy

Concept Drift Analysis

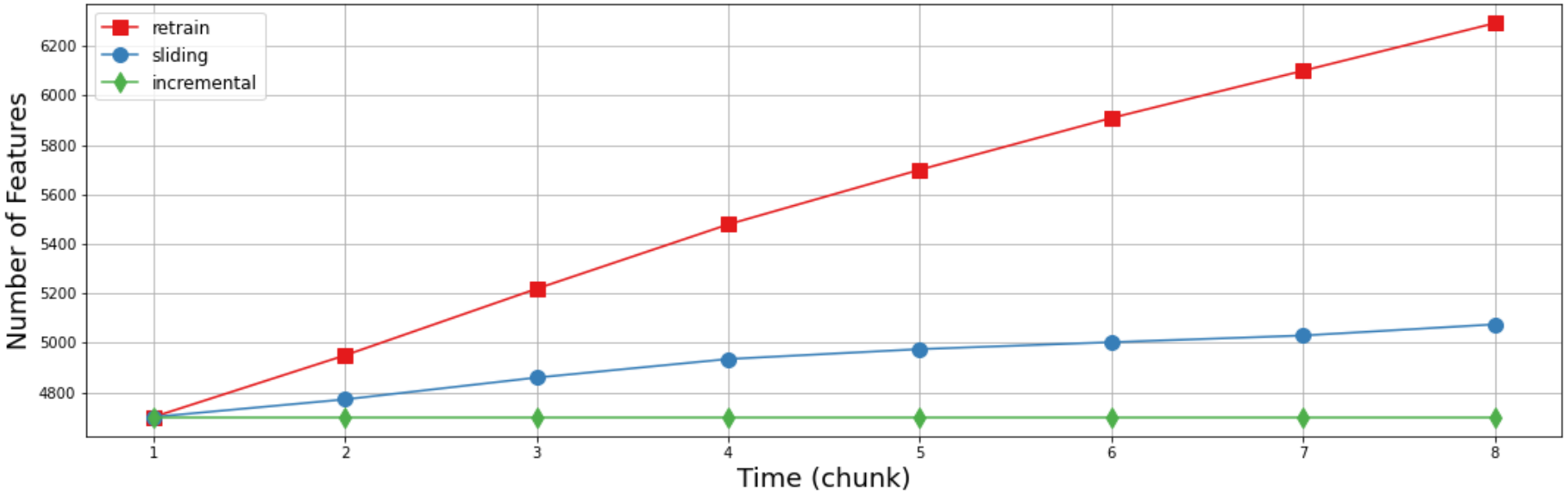


COMPARISON OF APPROACHES ALONG THE ONLINE MONITORING IN TERMS OF F_1^{PN} AND ACCURACY: AVERAGE VALUES AND AVERAGE RANKS.

	F_1^{PN}		Accuracy	
	Average	Avg. Rank	Average	Avg. Rank
Retrain	**0.6812**	**1.94**	**0.6763**	**2.25**
Sliding	0.6768	2.44	0.6744	2.31
Incremental	0.6758	2.19	0.6731	**2.25**
Ensemble	0.6358	4.13	0.6300	3.88
Static	0.6223	4.31	0.6144	4.31

- **Static** approach leads, in general, to the worst performance
  - a proper strategy for counteracting concept drift needs to be adopted
- **Ensemble** approach is still not enough to mitigate the impact of concept drift
- **Retrain**, **Sliding** and **Incremental** approaches perform comparably and they consistently outperform ensemble and static approaches.

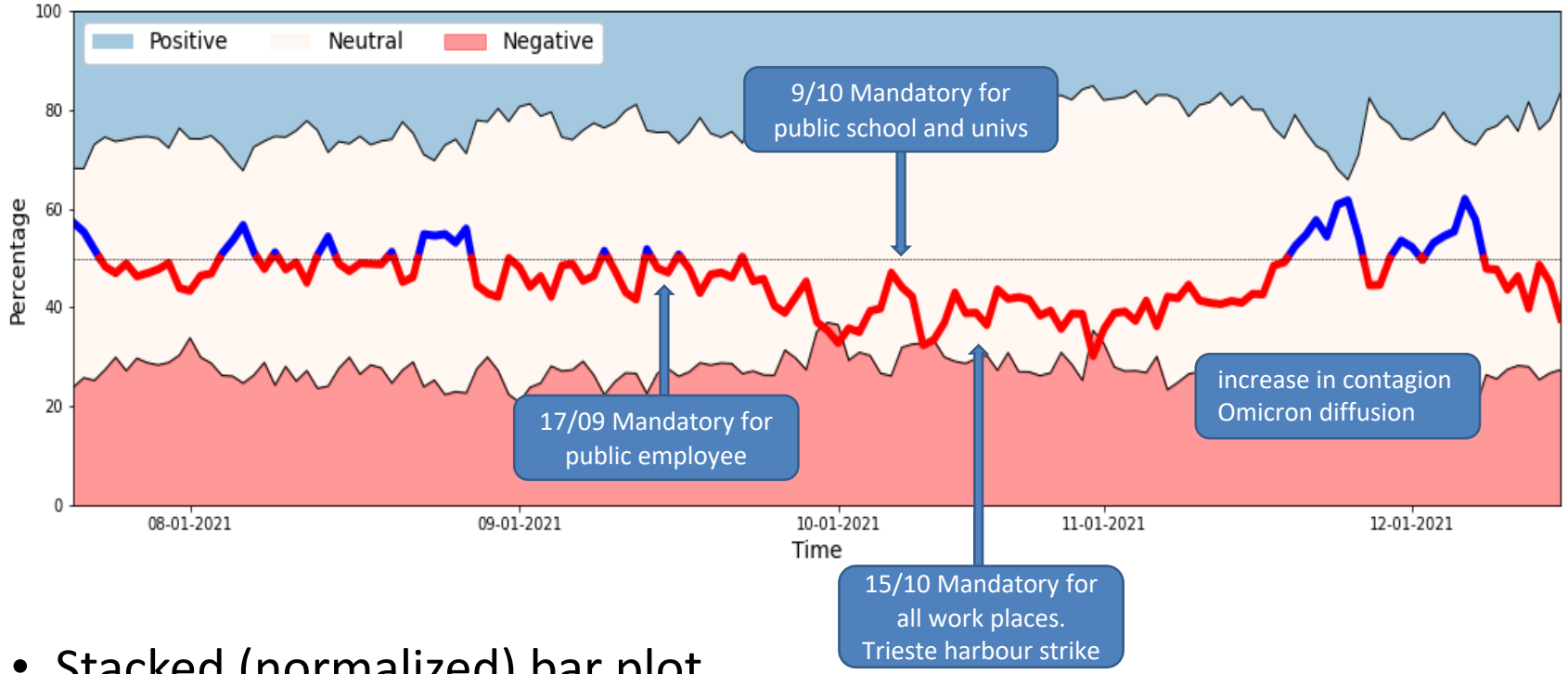
Concept Drift Analysis



- The **number of features** considered by the classifier at each chunk
  - grows significantly in the **Retrain** approach, as new tweets are added to the training set
  - grows moderately in the **Sliding** approach, as besides adding tweets we also discard outdated ones
  - remains constant in the **Incremental** case, since only the Bayesian classification model is updated

Selected for retrospective analysis

Stance of Italian Twitter users towards Green Pass.



Daily Stance formulation

$$Stance_i = \frac{Positive_i}{Positive_i + Negative_i}$$

- Stacked (normalized) bar plot
- *overall stance* (line in the central part of the graph)
  - Blue → biased toward positive stance (in favor of Green Pass)
  - Red → biased toward negative stance (against Green Pass).

References

E. D'Andrea, P. Ducange, A. Bechini, A. Renda & F. Marcelloni
Monitoring the Public Opinion about the Vaccination Topic from Tweets Analysis.
Expert Systems with Applications, 116:209–226, (2019)

A. Bechini, P. Ducange, F. Marcelloni & A. Renda
Stance Analysis of Twitter Users: the Case of the Vaccination Topic in Italy
IEEE Intelligent Systems, *36*(5), 131-139. (2020)

A. Bechini, A., Bondielli, A., Ducange, P., Marcelloni, F., & Renda, A.
Addressing event-driven concept drift in twitter stream: A stance detection application.
IEEE Access, *9*, 77758-77770. (2021)

Bondielli, A., Tortora, G. C., Ducange, P., Macri, A., Marcelloni, F., & Renda, A.
Online Monitoring of Stance from Tweets: The case of Green Pass in Italy.
In *2022 IEEE International Conference on Evolving and Adaptive Intelligent Systems*
(pp. 1-8). IEEE. (2022)