UN APPROCCIO BASATO SU DATI WEB REAL-TIME PER PREDIRE L'INFLUENZA

CARMELA COMITO, AGOSTINO FORESTIERO, CLARA PIZZUTI

CNR-ICAR

carmela.comito@icar.cnr.it

RESEARCH LINES

- **Early epidemics warning system**
  - **Mining Social Media Data to Predict, Detect and Track Disease Outbreaks**
    - *Influenza Prediction*
    - Real-time syndromic surveillance and early detection of emerging disease
    - Predicting spatio-temporal evolution of diseases with human mobility data
    - Diesease spread monitoring and modelization
- **Early disease diagnosis and treatment prediction**
  - **Clinical decision support in disease diagnosis and treatment**
    - Analysis and interpretation of radiology images
    - Treatment Impact prediction
      - Diagnostic and predictive models for therapeutic response of oncological and neurodegenerative diseases
      - Unsupervised pattern mining to predict the response to therapy and clinical outcome of the disease
      - Drug adversial reaction

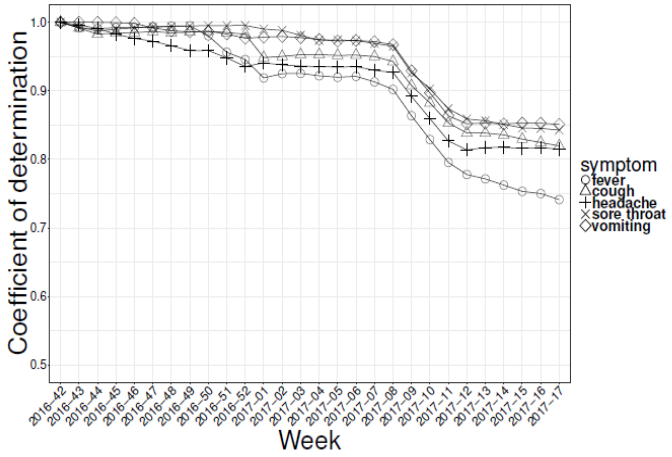
METHODOLOGIES AND TECHNIQUES

- Machine learning
  - Predictive models
  - Mining human mobility models
  - Trajectory pattern mining from GPS traces
  - Online clustering of social data streams
  - Real-time peak detection in tweet streams
  - AutoRegressive models
- Deep learning
  - Word Embedding
  - Long short-term memory (LSTM) recurrent neural networks (RNN)
  - Deep neural networks

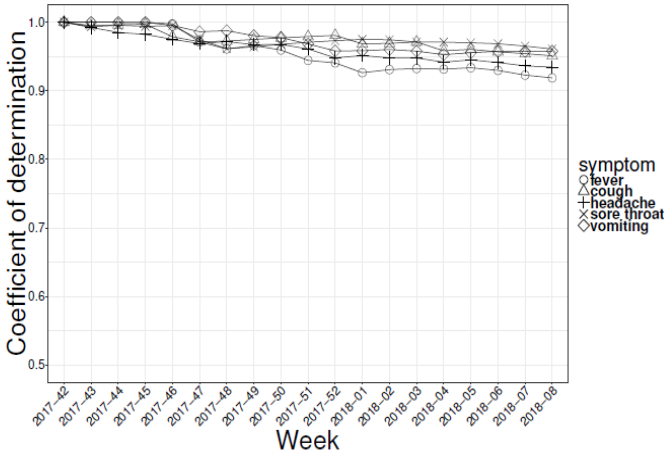
INFLUENZA PREDICTION - AUTOREGRESSIVE MODEL

$$y_{w+t} = \sum_{i=1}^{m} \alpha_i \mathring{y}_{w-i} + \sum_{j=0}^{l-1} \beta_j GT_{w-j} + \sum_{k=0}^{p-1} \gamma_k T_{w-k} + \delta$$

- y_w : predicted number of physicians visits due to ILI in week w
- $\mathring{y}_w$: real InfluNet data at week w
- GT_w : number of Google queries with ILI keywords at week w
- T_w number of unique Twitter users with ILI related tweets at week w
- t: number of weeks ahead we want to make the prediction.
- $GT_w = \sum_i r_{k_i} GT_{k_i,w}$ and $T_w = \sum_i r_{k_i} T_{k_i,w}$
- Pearson correlation $r_{k_i} = \frac{\sum_{w=1}^{n} (\mathring{y}_w - \bar{\mathring{y}})(GT_{k_i,w} - \bar{GT}_{k_i})}{\sqrt{\sum_{w=1}^{n} (\mathring{y}_w - \bar{\mathring{y}})^2} \sqrt{\sum_{w=1}^{n} (GT_{k_i,w} - \bar{GT}_{k_i})^2}}$



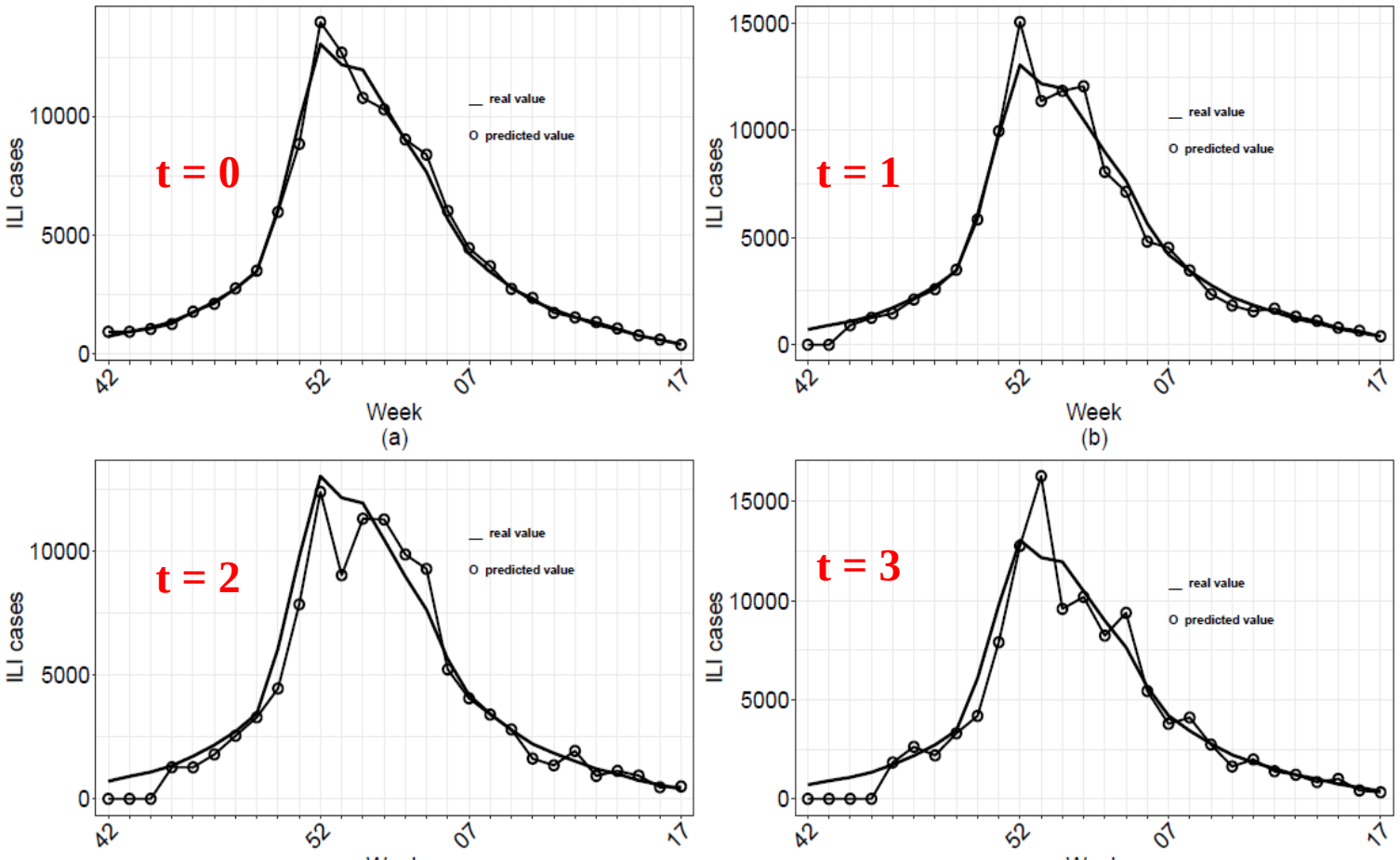
(a) Season 2016-2017



(b) Season 2017-2018

RESULTS: PREDICTED VS REAL VALUES

- Predictive model vs ground truth influenza data across time horizon t
  - For t=0 the predicted values follow very strictly the real values
  - With larger week lead (t) the forecasting model is less accurate
    - albeit still exhibiting good predictions



RESULTS: COMPARISON TO RELATED WORKS

- Real-time influenza forecast is improved by integrating surveillance data with Web-based social data
  - ***Prediction error* reduced up to 47% (RMSE)**
  - ***Pearson's correlation increased up to 24%***
- Only web-social data either only real-time (B_2) or also historic (B_3) is not enough to predict the ILI rate
  - high prediction error and low correlation
  - less accurate prediction than
    - only traditional surveillance data (B_1)
    - (even more) the proposed model M_4.

PERFORMANCE METRICS OF THE FORECASTING MODEL M_4 ACROSS t WITH RESPECT TO BASELINE MODEL

	Model	RMSE (%ILI)	MAE (%ILI)	MAPE (%)	r (corr)	R^2
t = 0	B_1	0.2456	0.1868	18.56	0.9763	0.9761
	B_2	0.3213	0.2822	29.12	0.8012	0.8601
	B_3	0.2619	0.2131	25.22	0.9622	0.9652
	B_4	0.2893	0.2723	26.67	0.9101	0.9584
	B_5	0.2914	0.2711	27.28	0.9063	0.9577
	M_4	**0.1972**	**0.1191**	**10.75**	**0.9867**	**0.9704**
t = 1	B_1	0.6418	0.3272	35.07	0.9465	0.8242
	B_2	0.9377	0.4421	47.31	0.7398	0.7113
	B_3	0.6942	0.3359	37.19	0.9486	0.8159
	B_4	0.7421	0.3561	38.17	0.9342	0.8288
	B_5	0.7552	0.3533	39.22	0.9221	0.8011
	M_4	**0.4080**	**0.2179**	**24.42**	**0.9623**	**0.9126**
t = 2	B_1	0.8069	0.6234	54.94	0.8875	0.7743
	B_2	0.9921	0.7377	68.98	0.7199	0.6401
	B_3	0.8292	0.6664	57.13	0.8345	0.7433
	B_4	0.8664	0.6781	59.39	0.8119	0.7412
	B_5	0.8598	0.6878	60.99	0.8064	0.7332
	M_4	**0.7020**	**0.5132**	**44.73**	**0.9212**	**0.8666**
t = 3	B_1	0.9144	0.6747	74.19	0.6192	0.6623
	B_2	1.3401	0.5992	63.98	0.587	0.5506
	B_3	0.9489	0.8171	87.95	0.6777	0.6263
	B_4	0.9851	1.7236	78.44	0.6358	0.5817
	B_5	0.9632	0.7327	80.09	0.62103	0.5976
	M_4	**0.7222**	**0.5824**	**62.53**	**0.8617**	**0.7855**
t = 4	B_1	1.6971	1.458	87.88	0.5663	0.5563
	B_2	2.8295	2.3171	98.15	0.5227	0.4761
	B_3	1.7111	1.7107	89.77	0.6173	0.5264
	B_4	2.4541	1.9208	90.28	0.5833	0.5024
	B_5	2.5487	1.9308	91.92	0.5963	0.5343
	M_4	**0.8972**	**0.6915**	**70.75**	**0.7912**	**0.6812**

CONCLUSIONS

- Investigated the use of web-based social data, namely, Google Trends and Twitter to track and predict influenza epidemic situation in the real world.
  - Results shown that the number of flu related tweets are highly correlated with ILI activity in Influnet data with a Pearson correlation coefficient of 0.9846.
- Built auto-regression models to predict number of ILI cases in a population as percentage of visits to physicians in successive weeks.
  - Tested our regressive models with the historic Influnet data and verified that Twitter data substantially improves our model's accuracy in predicting ILI cases.
- In view of the lag inherent in InfluNet's ILI reports, Web-based social data provides near real time assessment of influenza activity and can be used to effectively predict current ILI activity levels.
- *Opportunity to significantly enhance public health preparedness among the masses for influenza epidemic and other large scale pandemic.*