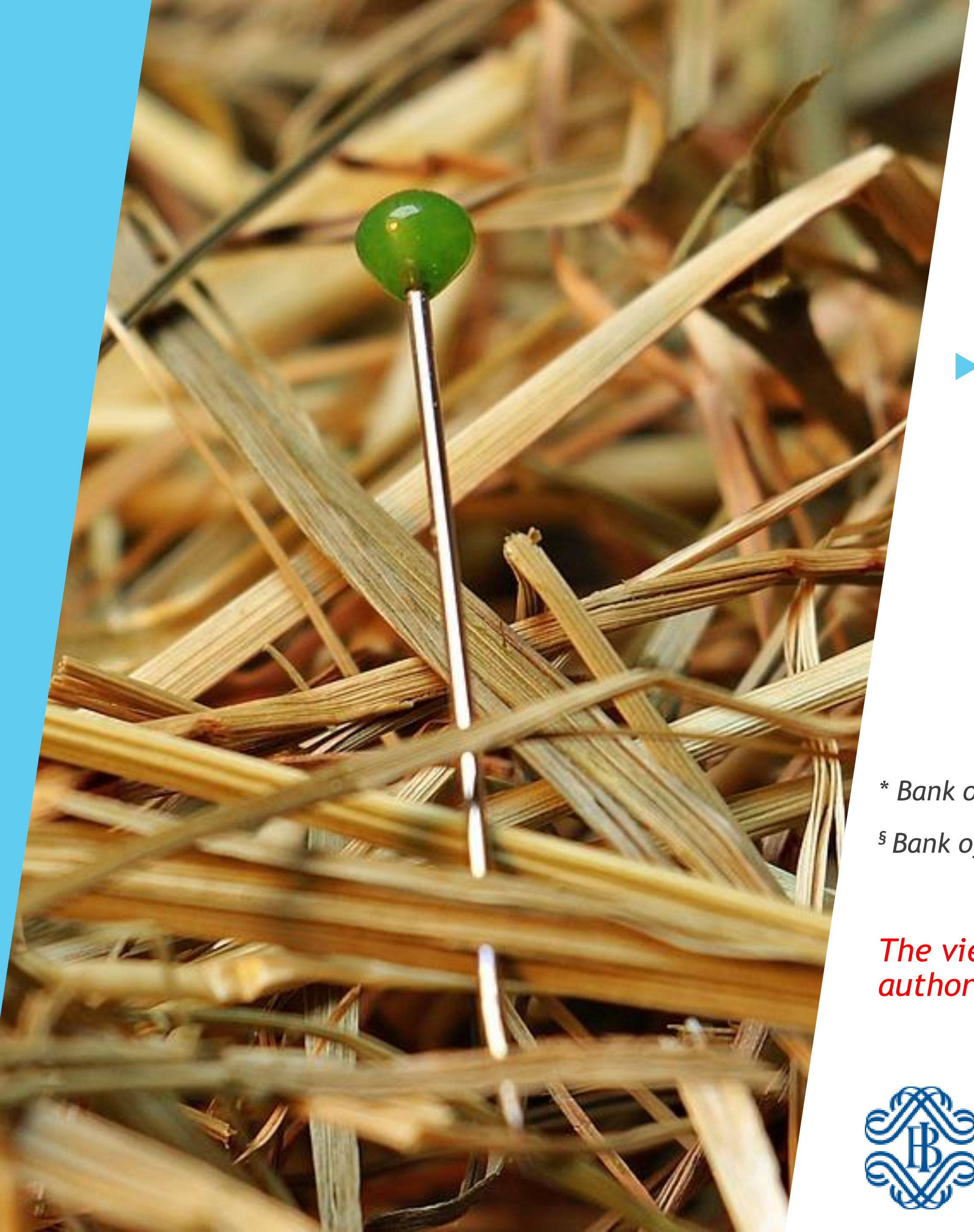


Deriving indicators from a large corpus of Italian documents

- By Marta Bernardini*, Pasquale Cariello*, Marco De Leonardis*, Juri Marcucci§, Filippo Quarta*, Alex Tagliabracci§

Workshop
«Big data & Machine Learning Applications for Central banks»

Rome, 21/10/2019

** Bank of Italy, DG for Information Technology*

§ Bank of Italy, DG for Economics, Statistics, and Research

BANCA D'ITALIA
EUROSISTEMA

Some example of TM in Bank of Italy

Type of Text	Docs' Length	Frequency	Main challenges
Tweets	Very short	Very High	Merging criteria; informal language; special characters/emoticons; noise;
real-estate dwellings	Short	Low	Duplicates; incongruences; images
Web site's scraping	Medium	Medium	Identify elements; structure's changes; access limitations.
Institutional Reports and Speeches	High	Low	Differences between sources; difficulty to obtain text (especially in the past)
Newspapers articles	High	High	Large corpus; performance issues; heterogeneity; spurious text;

Agenda

- How we build the corpus
- The tools we used
- The text mining pipeline
- Preliminary Results
- Conclusion & future works

ID='LAREP00020190501ef510002p'

la Repubblica

01-MAG-2019 — PUBLICATION DATE
Pagina 6
Foglio 1

I numeri dell'economia

Non è più recessione ma per Pil e lavoro solo una miniripresa

TITLE

**Nel primo trimestre l'Italia cresce dello 0,2%, la zona euro fa +0,4%
Merito soprattutto dell'export. In marzo giù al 10,2% i disoccupati**

SNIPPET

BYLINE

ROBERTO RHO, MILANO

Dopo due trimestri di crescita sottozero l'Italia esce dalla recessione tecnica. Un po' a sorpresa l'Istat certifica per il primo quarto dell'anno in corso un progresso del Pil dello 0,2% rispetto all'ultimo trimestre del 2018 e dello 0,1 per cento su base annua. Letta insieme al dato sulla disoccupazione, in calo al 10,2%, e al piccolo balzo della produzione industriale in gennaio e febbraio, l'appena percettibile inversione del Pil basta a cambiare un po' l'umore che si respira intorno all'azienda Italia. Cavalcano i dati dell'Istat, per primi, i rappresentanti del governo giallo- si è in costante ridimensionamento, in Italia è cresciuto in media di 1,5 punti all'anno negli ultimi cinque, principalmente a causa della debolezza della crescita). Ma anche perché il più 0,2 per cento italiano nel primo trimestre si confronta con il più 0,3 della Francia (più 1,1% anno su anno), con il più 0,7% della Spagna e con una stima di più 0,4% della media dell'Eurozona. La crescita italiana, dunque, è dimezzata rispetto a quella dei nostri partner europei. E ancora: il piccolo progresso del primo trimestre è frutto delle esportazioni, mentre la domanda interna (al lordo delle scorte) resta negativa. e quella giovanile, in particolare, dal 31,8 al 30,2%. Tutto merito dei 60 mila nuovi posti creati, 44mila dei quali a tempo indeterminato, che parrebbero frutto delle stabilizzazioni di alcune migliaia di contratti spinte dal carburante degli incentivi. Il tasso di occupazione, fa notare la Direzione Studi e ricerche di Intesa Sanpaolo, è salito al 58,9%, record almeno dal 2004, il che lascia prevedere «che le stime contenute nel Def sul tasso di disoccupazione, visto in salita quest'anno all'11%, siano eccessivamente pessimistiche».

BODY

TEXT=TITLE+SNIPPET+BODY

Query

PERIOD

Sub-query 1
«Economics & Finance»

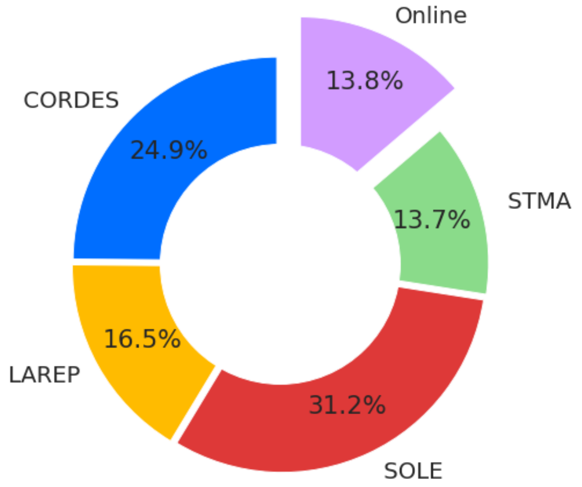
Sub-query 2
«Inflation»

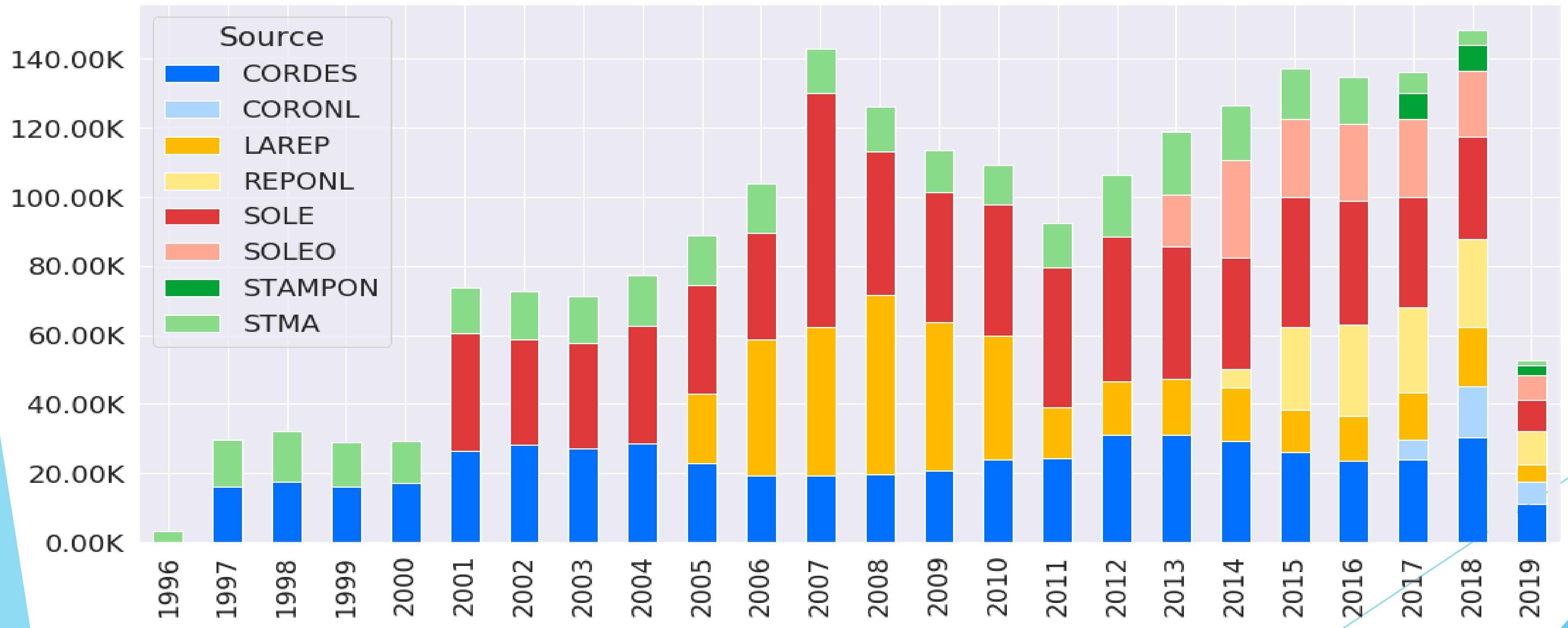
```
econ_final_v3: from 01/01/1995 to 05/05/2019, (((economi*
or finanza or finanziar* or tass* or finanze or moneta
or monetari* or "banca centrale" or BCE or bankit* or
"banca d'italia" or ns=(e12 or ecat or mcat or ccat)
or prezzo or prezzi or "costo della vita" or inflaz*
or "caro bollette" or "caro prezzi" or "caroprezzi" or
"benzina alle stelle" or "bolletta salata" or "caro
affitti" or "caro benzina" or "caro carburante" or
"caro gas" or deflaz* or disinflaz* or ribass* or
"meno caro" or "bollette pi leggere" or salar* or
stipend* ) not ( (ns=gspo) or (ns=gent) or
(ns=gwere))) and (rst=cordes or rst=coronl or rst=stma
or rst=stampon or rst=sole or rst=soleo or rst=larep
or rst=reponl)) and (la=It)
```

Source & Language

documents: 2,158,637
words: 566,349,655

Query


Online
13.8%
CORDES
24.9%
STMA
13.7%
LAREP
16.5%
31.2%
SOLE


Source
CORDES
CORONL
LAREP
REPONL
SOLE
SOLEO
STAMPON
STMA
140.00K
120.00K
100.00K
80.00K
60.00K
40.00K
20.00K
0.00K
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

Our Approach

Original Articles

Words are the New Numbers: A Newsy Coincident Index of the Business Cycle

Leif Anders Thorsrud

Received 01 Nov 2017, Accepted author version posted online: 09 Aug 2018, Published online: 05 Nov 2018

Download citation https://doi.org/10.1080/07350015.2018.1506344 Check for updates

Transparency and deliberation within the FOMC: a computational linguistics approach

Stephen Hansen, Michael McMahon, Andrea Prat · Published 2014 · DOI: 10.1093/qje/qjx045

We follow current literature practices [Thorsrud (2018); Hansen et al. (2014), with some differences:

- Italian Language (no English Translation)
- Ad Hoc Query (no full newspaper)
- Different methods for selecting words (dictionary)
- Computational challenges
- In-sample & Out-of-sample validation (Inflation forecasts)

Tools and Languages

Big Data platform on-premise based on Hadoop

- Distributed/Parallel computations;
- In-memory computations;
- Spark MLLib
  - Number of executors: 8
  - Number of executor cores: 4
  - Executor Memory: 20G
  - Driver Memory: 10G
- Python language (Jupyter Notebooks);

~ 2 hr.

jupyter 05_ECON_FINAL_V3_TopicAnalysis_Analyze_Choice_K_LDA_Q2_Zipf Last Checkpoint: Last Tuesday at 8:49 AM (autosaved)

Control Panel Logout

File Edit View Insert Cell Kernel Help

Python 2.7 (Spark2)

Code CellToolbar

Optimal Number K of topics for LDA

In [5]:

```
appended_data = []
#for testToEvalute in testToEvaluteList:
pathSingleFile = evaluation
dfSingleFile = pd.read_csv(pathSingleFile, sep=';').sort_values(['K'])
appended_data.append(dfSingleFile)

dfPandasFileTest= pd.concat(appended_data)
#dfPandasFileTest = pd.read_csv('/home/m030089/Factiva/evaluation_models_'+str(testToEvalute)+'.csv', sep=';').sort_values(['K'])
```

In [7]:

```
dfPandasFileTest[['K','Train_logPerplexity','Test_logPerplexity']]
```

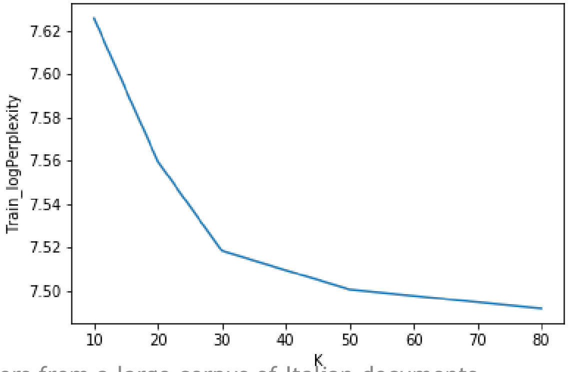
Out[7]:

	K	Train_logPerplexity	Test_logPerplexity
0	10	7.625612	7.641212
1	20	7.559505	7.592200
2	30	7.518250	7.567824
3	50	7.500432	7.583376
4	80	7.491650	7.615186

In [8]:

```
#LogLikelihood
%matplotlib inline
sns.lineplot(x='K', y='Train_logPerplexity', hue='Pipeline',
             data=dfPandasFileTest, legend=False)
```

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x77eba50>



Text Mining Pipeline

We worked only on 'Inflation' subquery, removing online sources (~300k docs)

Cleaning the corpus

- Lower case
- Remove numbers
- Remove punctuation
- Tokenization
- Filter
- Dictionary for LDA

Finding the topics

- LDA (Latent Dirichlet Allocation)
- Select best model

Constructing time series

- Daily topic intensity
- Sentiment Analysis
- Finding correlations
- Time series forecasting

Cleaning's phase

	Raw text	Unique words	Identify collocations	Remove stopwords	Stemming	TF-IDF adjustment	Min DF
Number of words	221,080,503	508,605	508,952	492,246	294,104	123,315	9,942

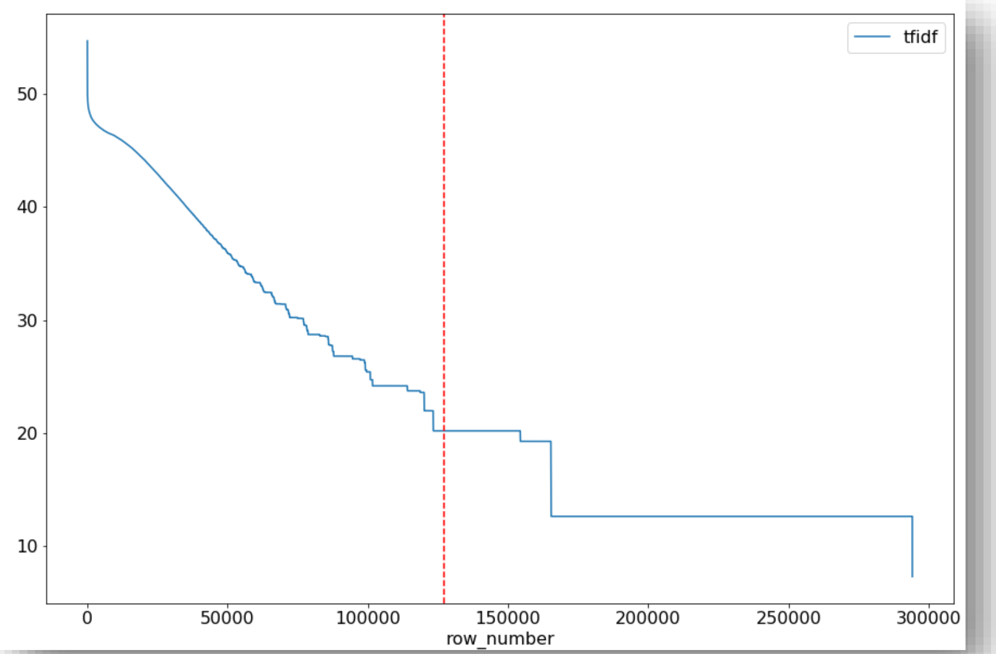
- Lower case
- Remove numbers
- Remove punctuation
- Tokenization (split words)
- Stopwords (Italian & English)
- Remove common names and surnames (from 'byline' field)
- Remove words with length < 3 or > 26
- Lemmatization/Stemming
- Bigram/Trigram
- TF-IDF or ZIPF's Law

Building the Dictionary

TF-IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

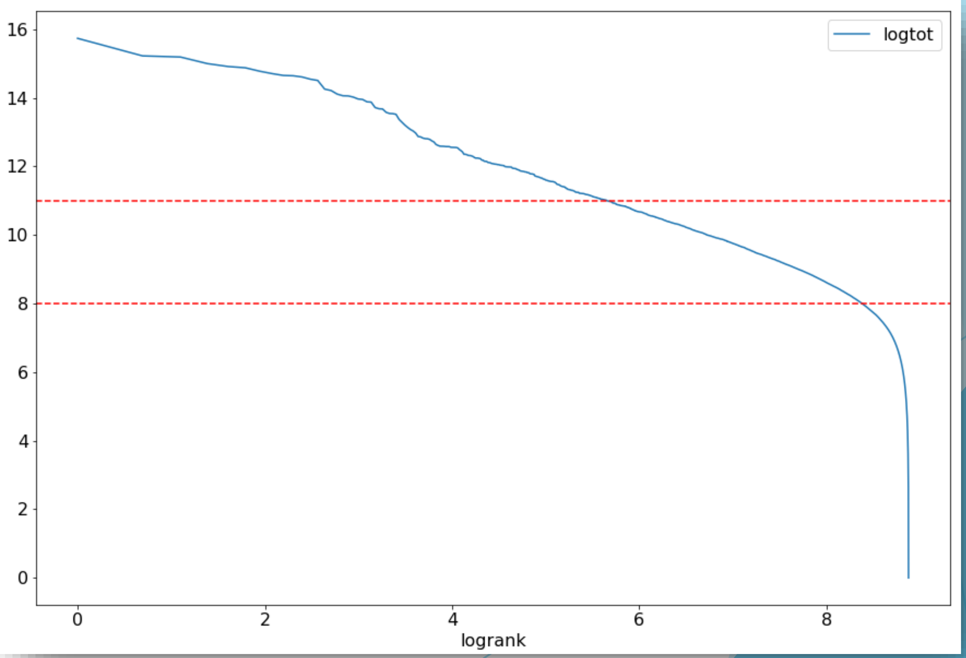
TF/IDF (25% totale; DF>300) → ~10.000 words



ZIPF's LAW

"the frequency of any word is inversely proportional to its rank in the frequency table"

Zipf Law (8 < log(freq(*w*)) < 11) → ~3.500 words

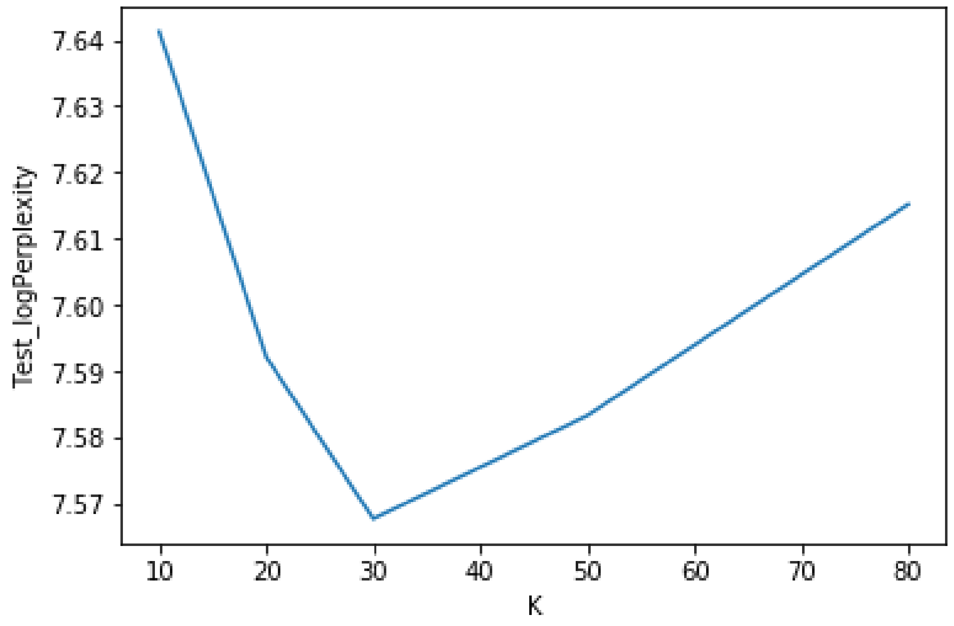


Latent Dirichlet Analysis

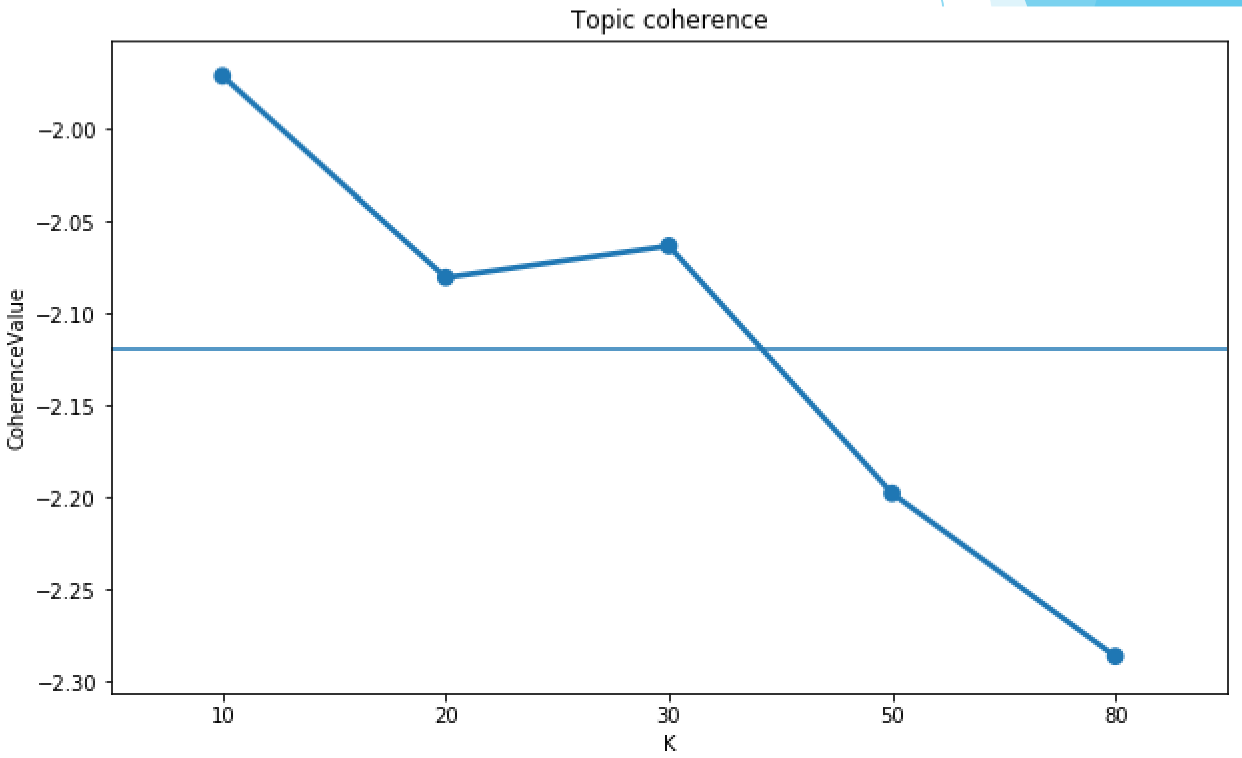
- Finding the LDA Model with best K (Number of Topics)
- List of K values used [10, 20, 30, 50, 80]
- TrainSet [80%] – TestSet [20%]
- Optimizer: Online variational Bayes
- Alpha (Doc-Concentration) = uniformly (1.0 / K) [default]
- Beta (Topic-Concentration) = (1.0 / K) [default]
- Evaluation Metrics:
  - Topic Perplexity *(how model captures the distribution of the held out set)*
  - Topic Coherence *(the degree of semantic similarity between its high scoring words)*

LDA metrics

Perplexity



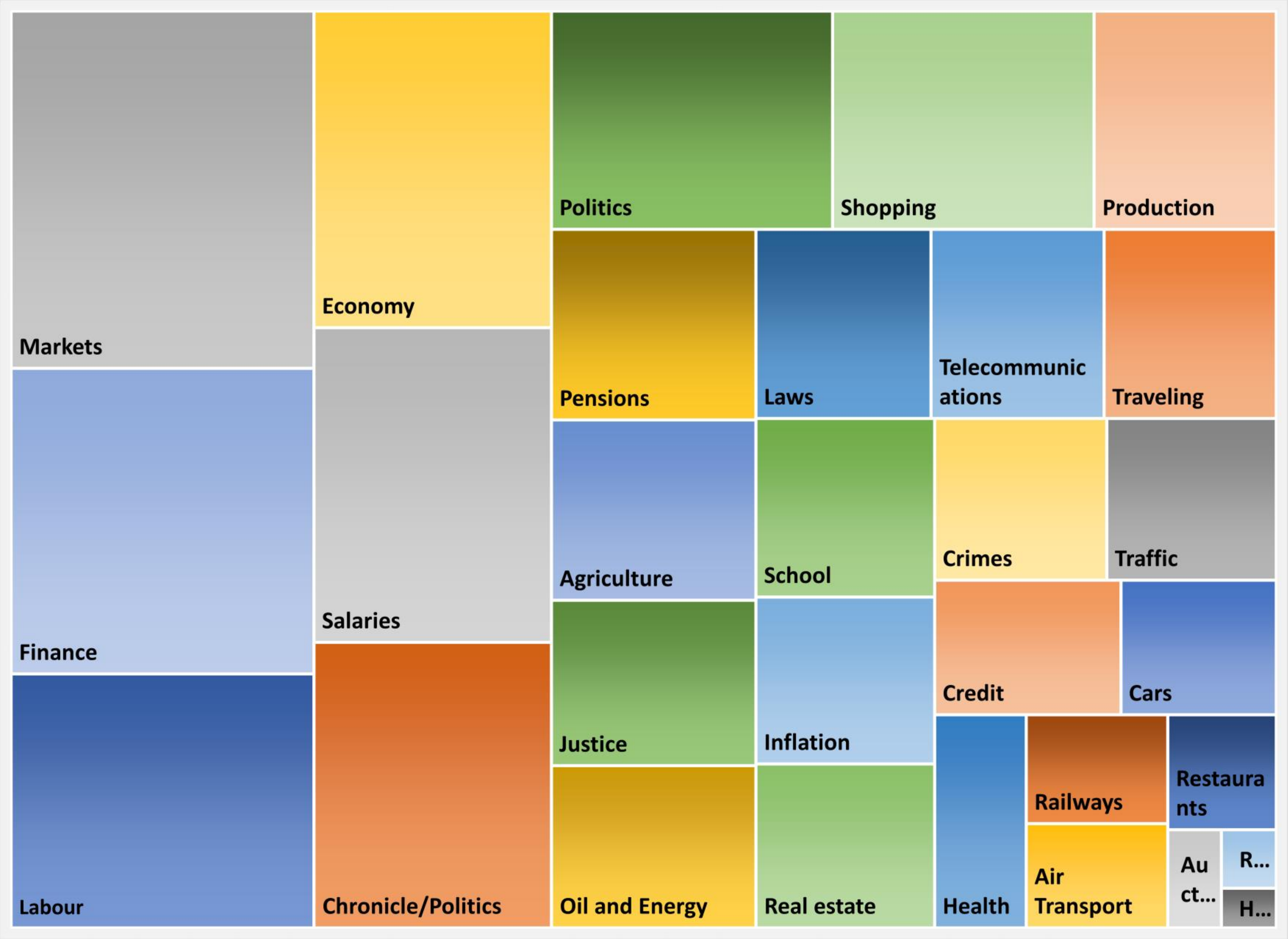
Coherence



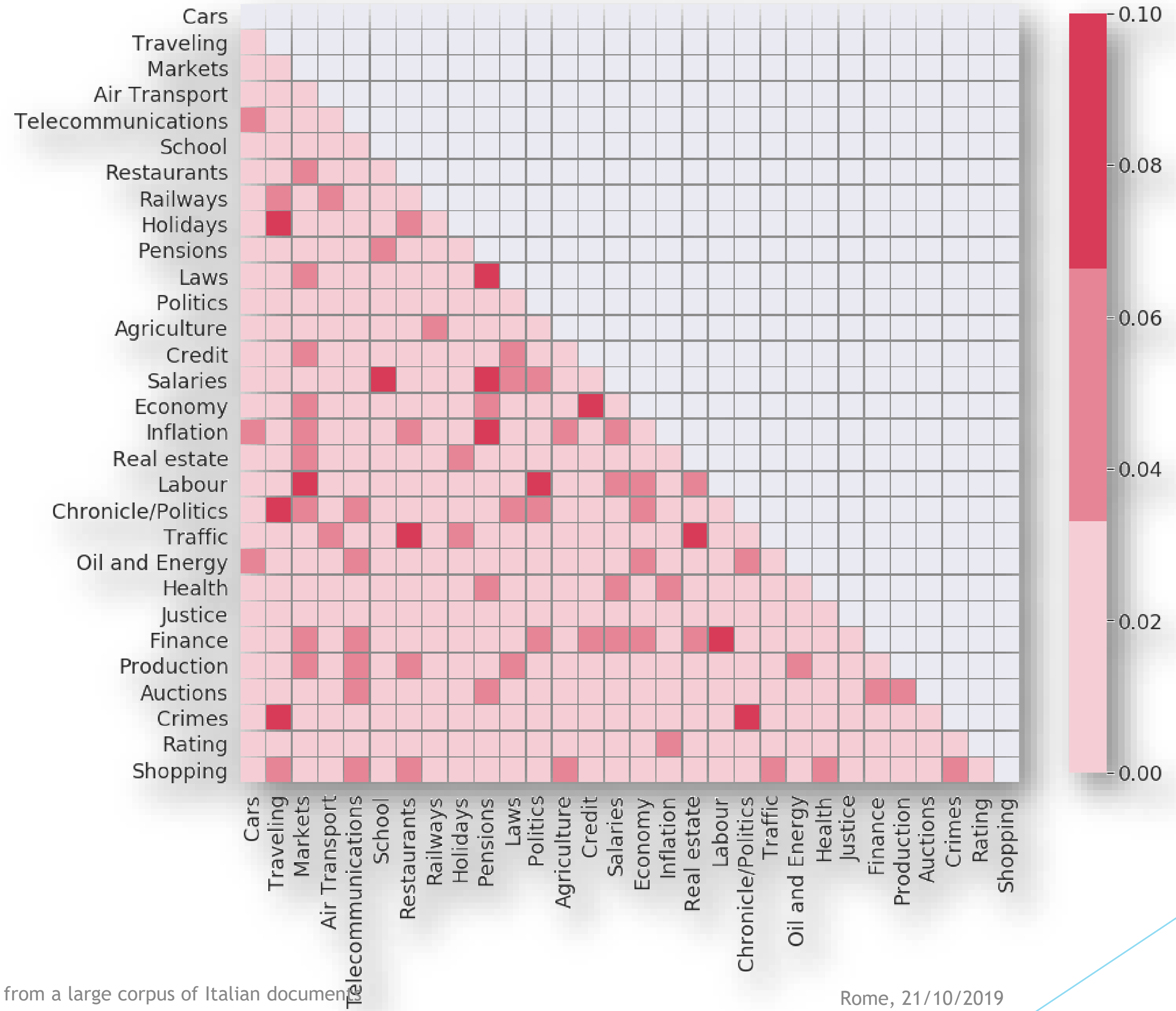
News Topics (K = 30)

topic	name	words
1	**Cars**	cars, version, engine, vehicles, gasoline, liter, volkswagen, chrysler, gamma, suv, diesel, launched, gear, automatic, electric
2	**Traveling**	children, kids, journey, tourists, people, family, holidays, to live, tells, night, hotel, reservations, help, hotels, take
3	**Markets**	dollars, rise, performance, discount, investment, sale, analysts, actions, wall_street, american, indicated, btp, loss, wait, be worth
4	**Air Transport**	company, plane, passengers, ships, low_cost, pilots, climbing, transport, malpensa, insurance, route, ryanair, traffic, air_france, wanted
5	**Telecommunications**	digital, customers, telephony, dollars, offer, smartphone, use, data, users, electronics, launched, software, function, technology, calls
6	**School**	students, study, teaching, training, salary, professional, university students, schools, graduates, temporary workers, graduate, employment, researchers, staff, employees
7	**Restaurants**	revenues, closed, ugly, turnover, disabled_access, drink, pizza, useful, margins, sale, dividends, open, positive, restaurant, menu
8	**Railways**	chinese, trains, transport, ticket, tons, steel, speed, journey, railway, railways, station, minutes, ilva, duties, siderurgy
9	**Holidays**	liguria, beach, restaurant, tourists, stability, season, shower, beach club, local, bath, concession, pool, beach umbrella, holidays, sand
10	**Pensions**	pension, article, decree, paragraph, intended, income, fiscal, tax, contributions, application, payment, employees, december, expense, within
11	**Laws**	article, paragraph, application, contract, decree, such, intended, subjects, obligation, law, indicated, relative, procedure, provision, rule
12	**Politics**	reform, premier, votes, parliament, left, elections, electoral, candidate, agreement, wants, voters, theme, opposition, senate, referendum
13	**Agriculture**	production, wine, milk, agricultural, quality, agriculture, tons, consumption, exports, kilo, breeders, harvest, supply chain, wheat, meat
14	**Credit**	credit, loans, debt, mortgages, banking, financing, liquidity, financial, institutions, loss, bankruptcy, obligation, rate, repayment, bad debts
15	**Salaries**	unions, employees, salaries, strike, salary, protest, blockade, regional, expense, contract, approved, municipalities, managers, announced, agreement
16	**Economy**	inflation, pil, debt, deficit, recession, measures, eurozone, too much, american, unemployment, world, fiscal, financial, monetary, expectations
17	**Inflation**	hundred, expense, gasoline, tariffs, income, taxes, fuels, average, inflation, price increases, cents, consumption, women, bills, growth
18	**Real estate**	real estate, inhabitants, rent, apartments, area, owners, building, property, housing, local, renovation, land, investment
19	**Labour**	contract, reform, productivity, unions, need, agreement, theme, represents, resources, intervention, performance, need, investment, necessary, measures
20	**Chronicle/Politics**	power, death, american, that, people, to live, perhaps, newspapers, parliament, man, to hear, remember, left, dollars, law
21	**Traffic**	parking, workers, hold, factory, area, open, traffic, local, tax, time, half, roads, closed, entrance, firm
22	**Oil and Energy**	energy, oil, electric, production, plants, dollars, nuclear, petroleum, emissions, medium, opec, supplies, world
23	**Health**	drugs, waste, sanitary, care, doctor, asl, hospitals, pharmaceutical, saipem, evil, good, medicines, regions, landfill, expense
24	**Justice**	prosecution, power of attorney, investigation, magistrates, crime, court, suspects, trial, justice, judicial, affair, conviction, corruption, legal, false
25	**Finance**	actions, offer, partners, participation, controls, opa, acquisition, agreement, merger, mediobanca, cda, transfer, holding, investment, financial
26	**Production**	registered, data, hundred, estimates, increase, previous, indicated, confirm, signals, grow, result, decrease, production, sign, positive
27	**Auctions**	auction, tomorrow, wednesday, organization, data, friday, thursday, october, participation, november, monday, tuesday, dedicated, february, december
28	**Crimes**	police, arrest, police, people, tells, finished, seizure, reported, drugs, death, agents, criminals, mafia, man, palestinian
29	**Rating/Investments**	information, rating, equity, various, redemption, flexibility, tariffs, balance sheets, guarantee, yield, subscription, index, daily, investment_grade, reserves
30	**Shopping**	shops, customers, balances, quality, idea, tells, good, think, buy, true, clothing, open, choice, better, search

Topics concentration



Topics – correlation



Time Series (Topics over time)

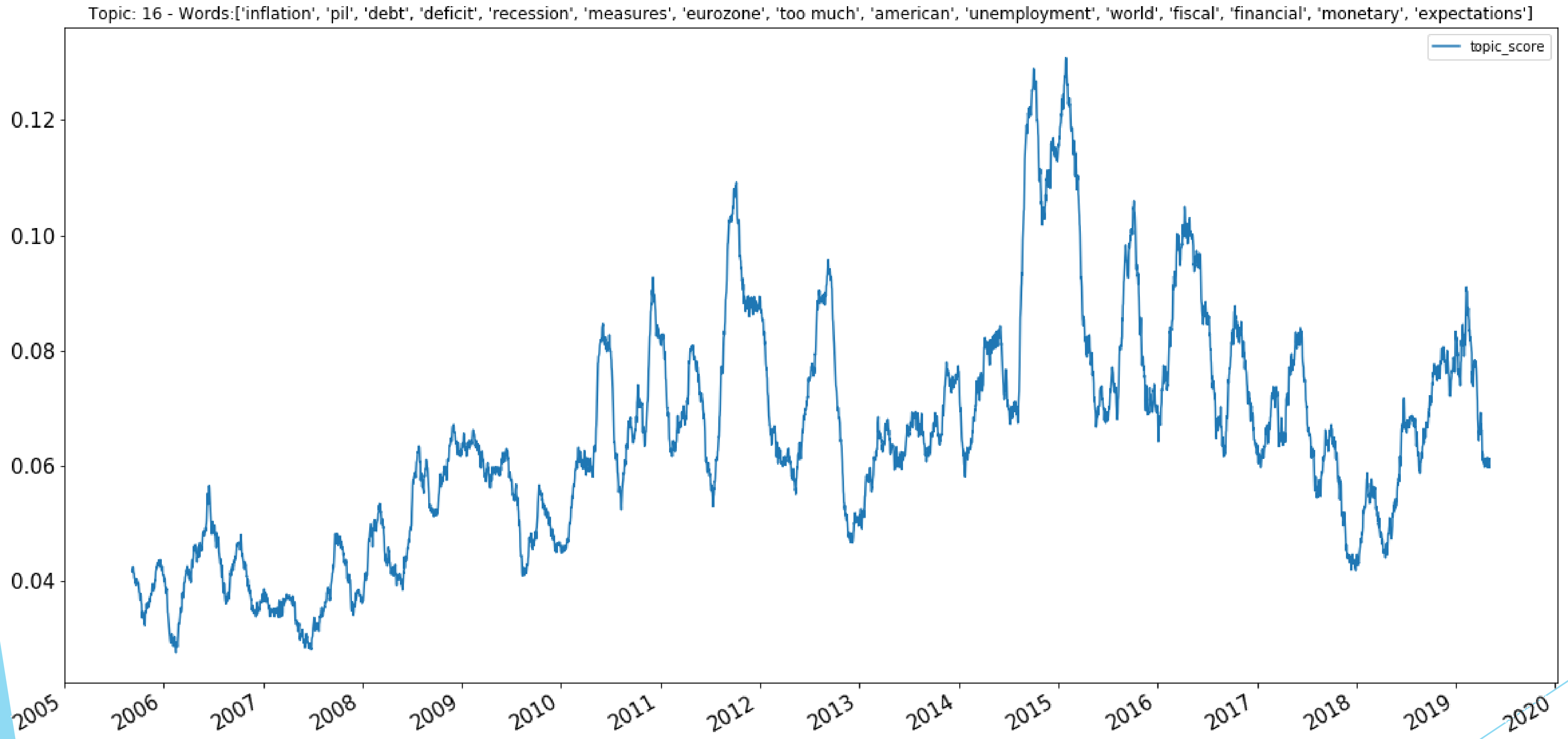
- **Daily frequencies**: we collapsed all the articles for a particular day into one document and then we computed, using the estimated word distribution for each topic, the topic frequencies for this newly formed document. This yields a set of K daily time series;

- **Sentiment analysis**:
  1. We adopt the Italian dictionary CNR to infer the number of negative and positive words for each article (https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-73?show=full#)
  2. Totally we have *25.098* words, but we keep *6.453* words: they are the words with strongest polarities (≤ -0.5 and ≥ 0.5)
  3. For each day and topic, find the article that is best explained by each topic, and from that identify the tone of the topic, that is, whether or not the news is positive or negative

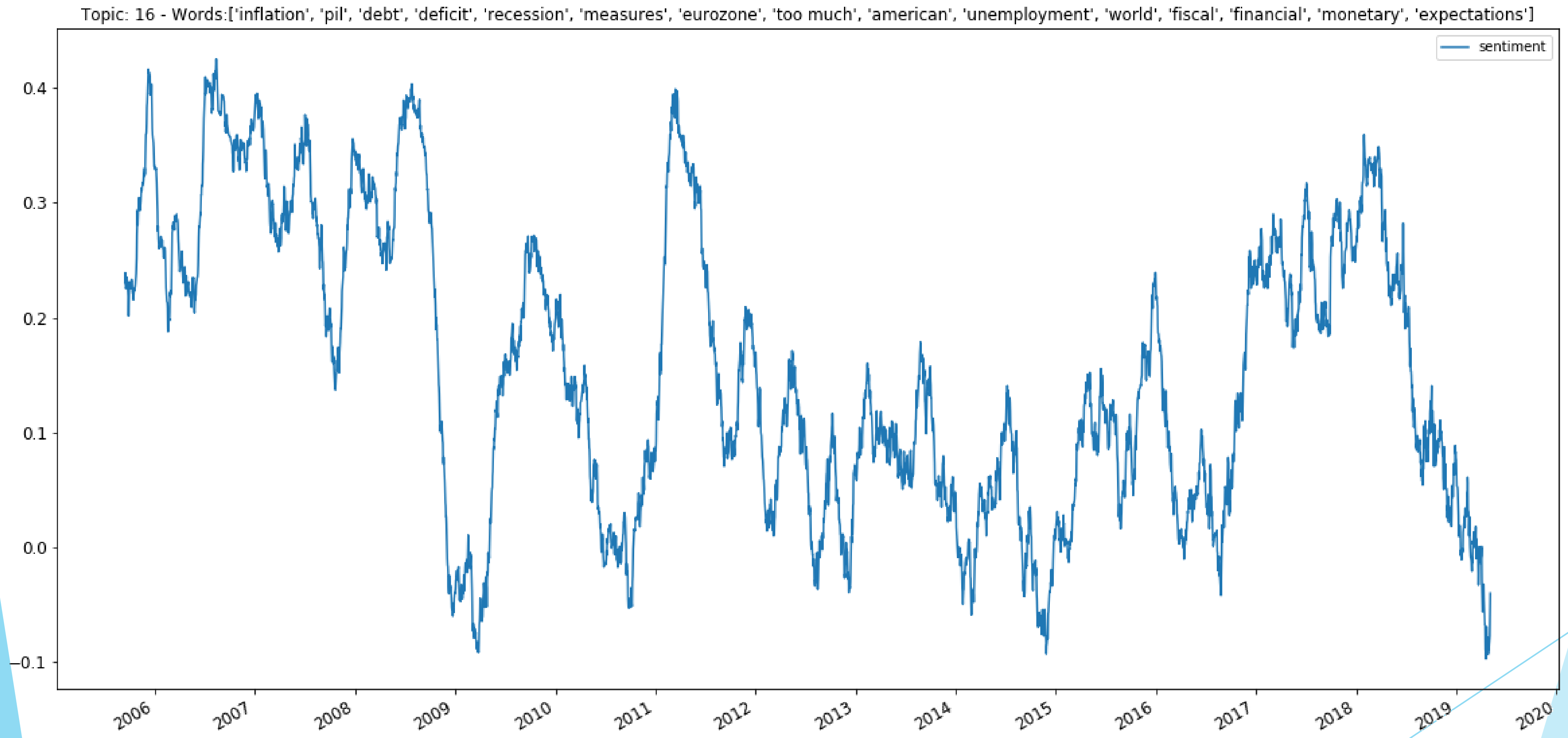
$$Pos_{t,m_t} = \frac{\#positivewordsm_{m_t}}{\#totalwords_{m_t}} \quad Neg_{t,m_t} = \frac{\#negativewords_{m_t}}{\#totalwords_{m_t}}$$

$$S_{t,m_t} = Pos_{t,m_t} - Neg_{t,m_t}$$

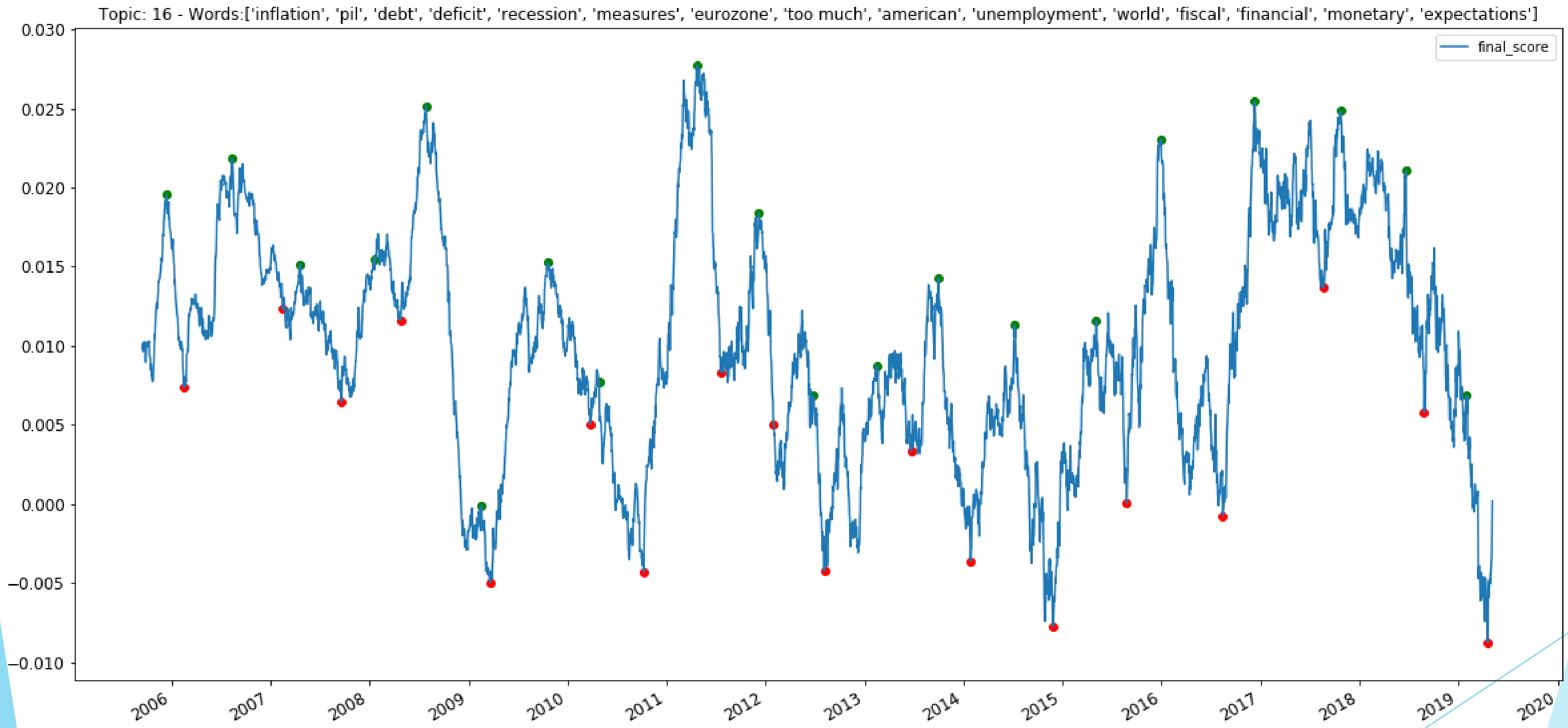
Topic 16 – Intensity



Topic 16 – Sentiment



Topic 16 – Intensity * Sentiment

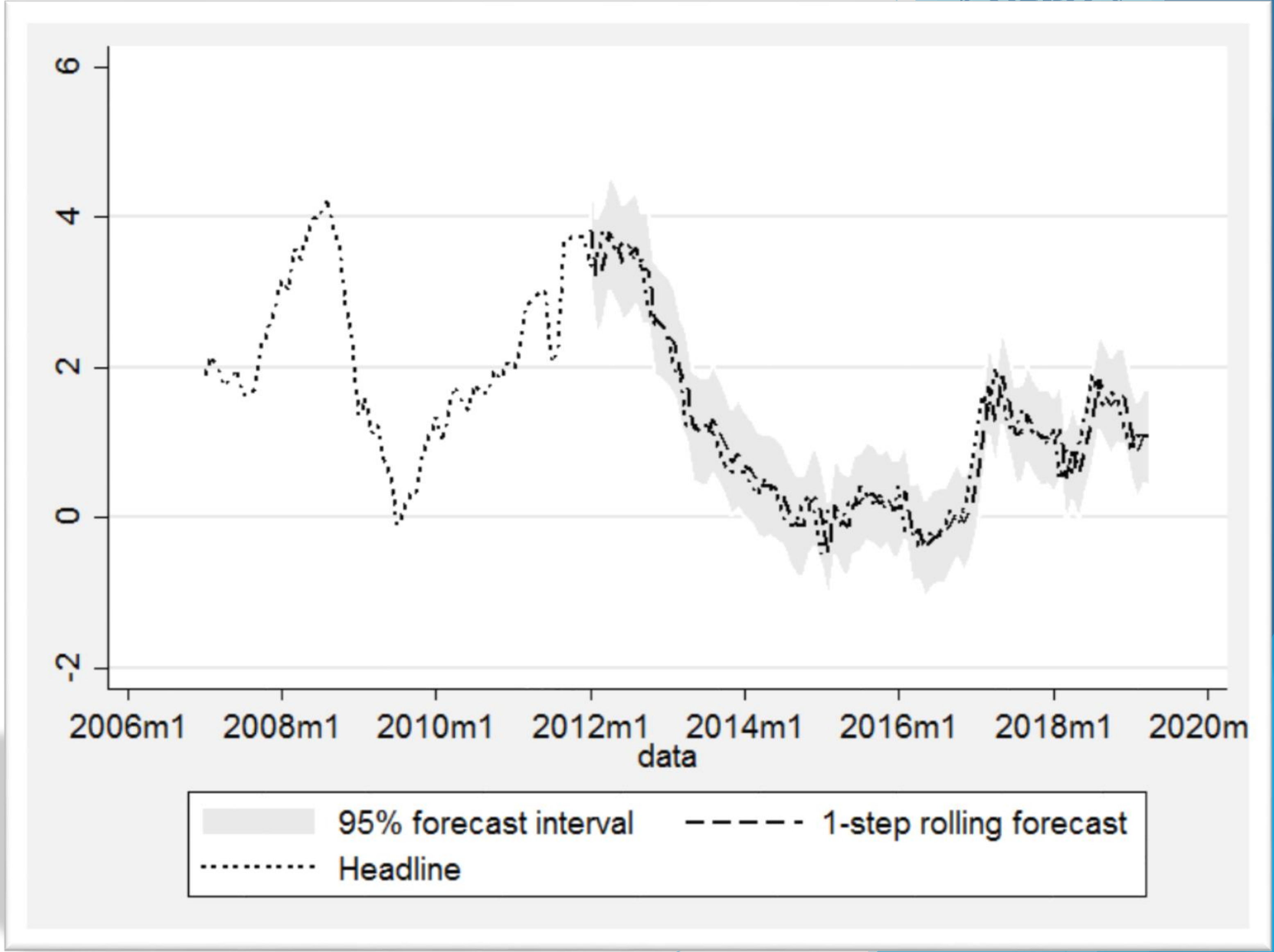


Forecasts

HICP (Harmonized Index of Consumer Prices) = Year-on-year rate of change at the monthly frequency.

Out-of-sample exercise on a rolling window of five years.

Forecast for Headline Inflation (HICP) and the relative prediction obtained with a AR(1)-X model (Benchmark).

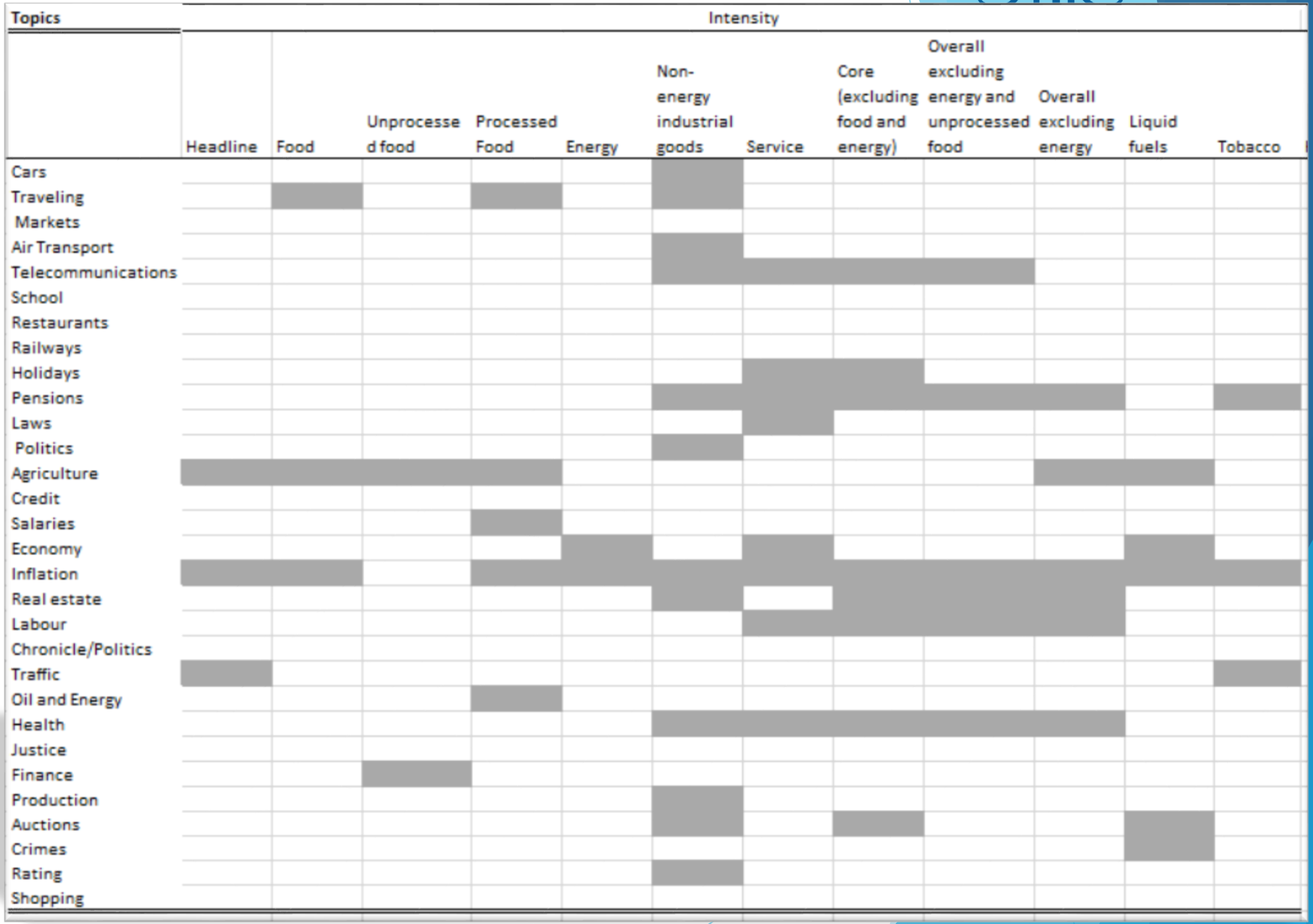


Forecasts

AR(1) vs. AR(1)-X (plus Intensity indicator for each topic).

Rolling window of five years (60 monthly observations).

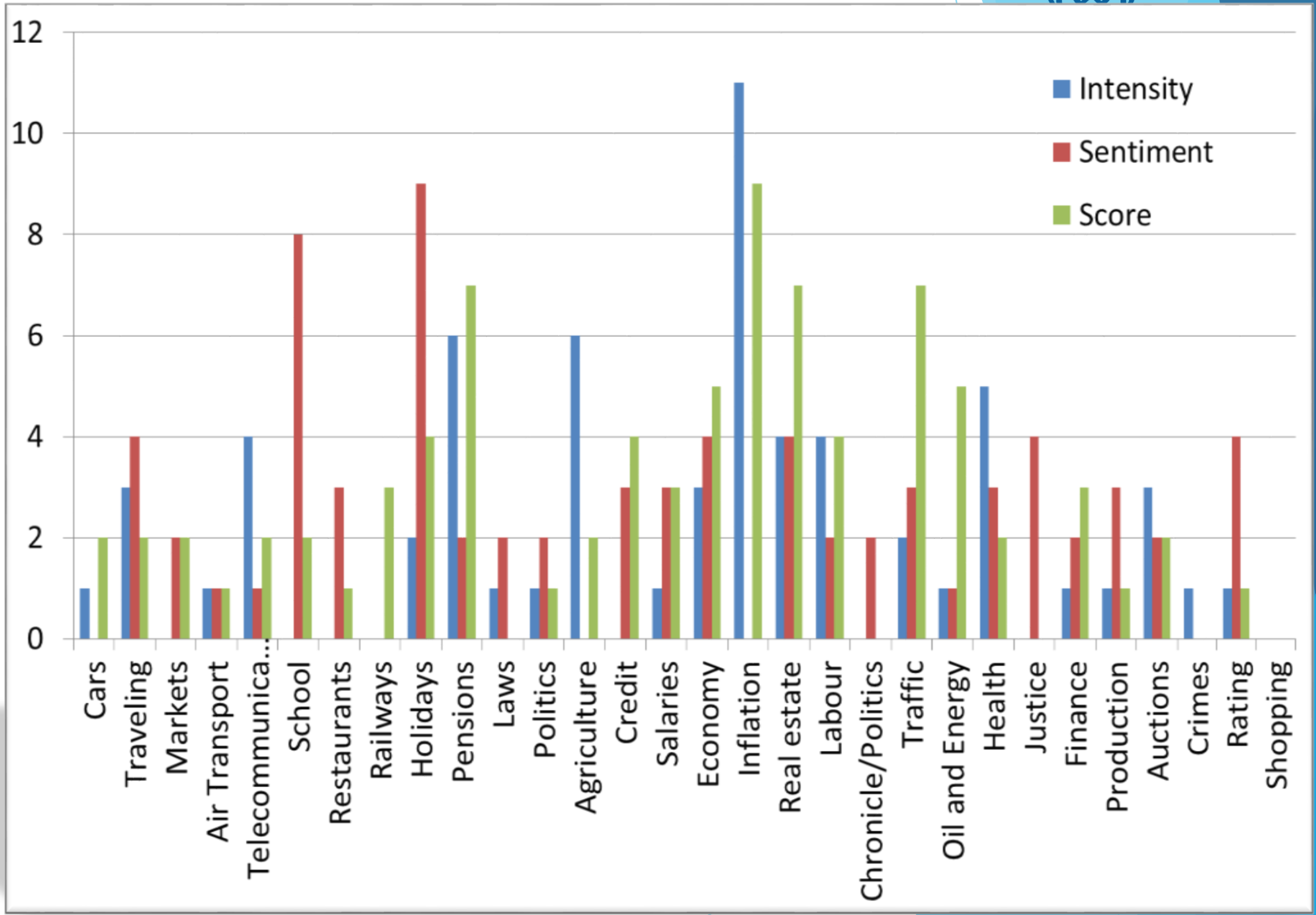
The grey cells indicate that the Root Mean Squared Error (RMSE) of the AR(1)-X is lower than that of the AR(1).



Predictive Power

Y-Axis: number of times each news topic on the x axis helps predicting the inflation rate with respect to an AR(1) benchmark using a rolling window of 60 months.

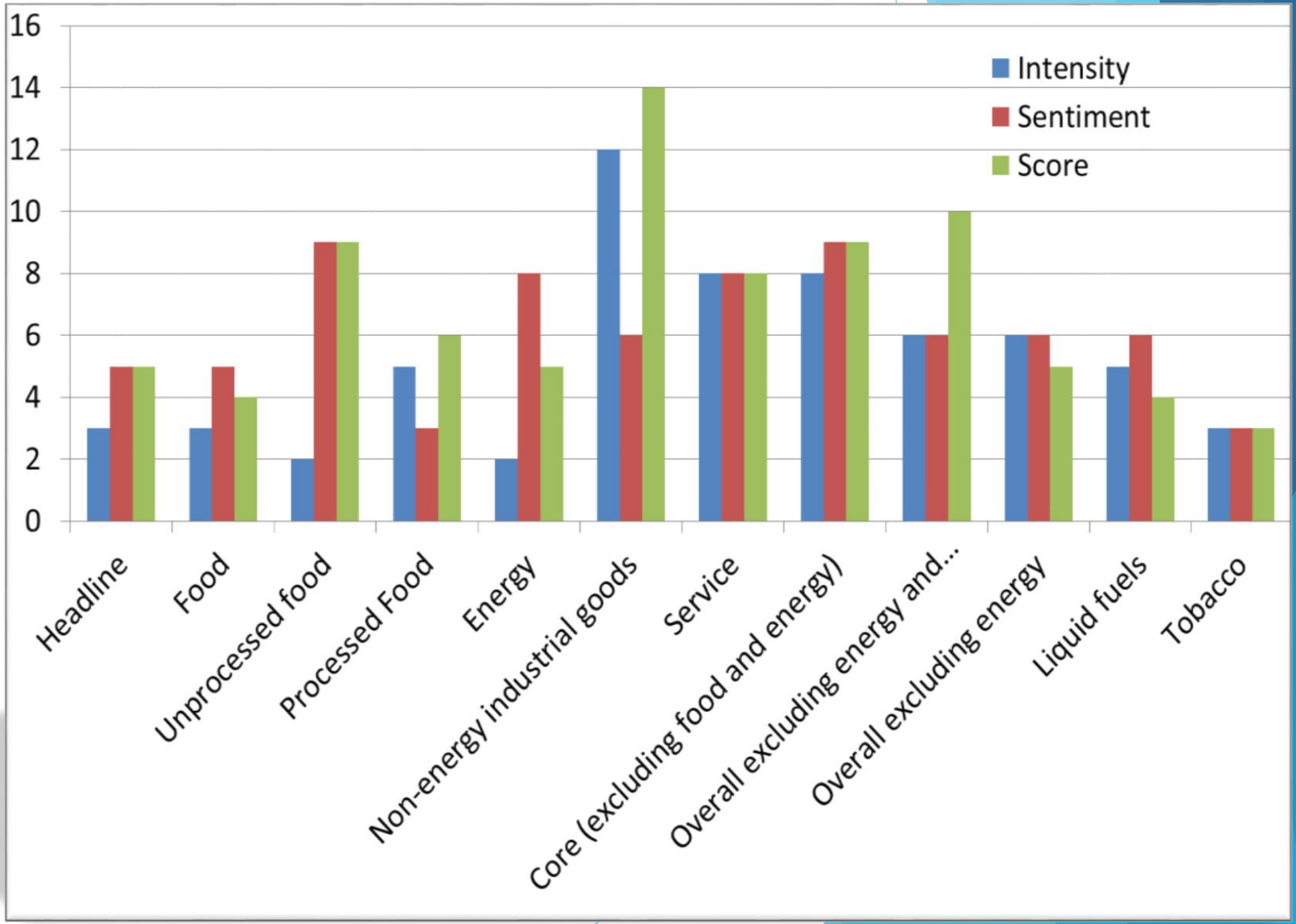
We depict a different bin for each different topic measure (intensity, sentiment and score, which is the intensity weighted for the score).



Predictive Power

Y-Axis: how many topic-based models outperform the benchmark using intensity, sentiment or the score of each topic.

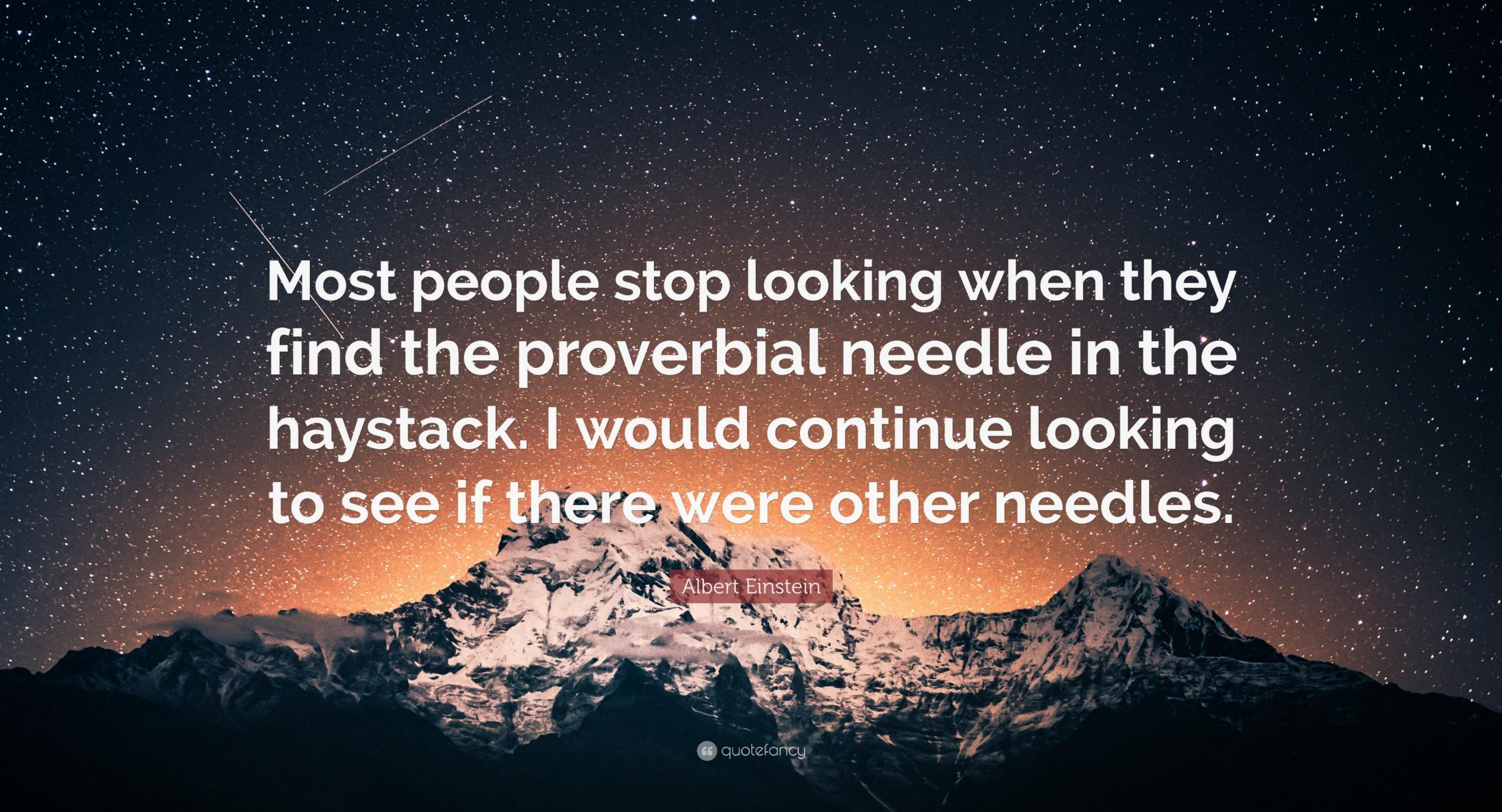
As highlighted in the previous literature for other languages - see for example (Thorsrud, 2018) – weighting the intensity indicators using sentiment or tonality does indeed help in predicting the variable of interest.



Conclusions and future work

- We build a large corpus of articles from Italian newspaper related to process and inflation from queries against Factiva archives;
- After filtering, we calculate 30 topics that exhibit good coherence , low correlation and uniform distribution of the articles;
- Indicators derived from that topics revealed some additional predictive power against a simple benchmark model in forecasting inflation.

- Further researches are already in progress, to improve the cleaning phase and to better quantify the informative gain from the news topics;
- We are also working on the other sub-query, concerning monetary policy and economic phenomena in general (~700k docs).


Most people stop looking when they find the proverbial needle in the haystack. I would continue looking to see if there were other needles.
Albert Einstein
quotefancy

Thank You

for your precious time and attention