The European Commission’s science and knowledge service

Joint Research Centre

European Commission

Monitoring the business cycle with fine-grained, aspect-based sentiment extraction from news

Luca Barbaglia, Sergio Consoli, Sebastiano Manzan

European Commission, Joint Research Centre, Directorate A-Strategy, Work Programme and Resources, Scientific Development Unit, Via E. Fermi 2749, I-21027 Ispra (VA), Italy.

{name.surname}@ec.europa.eu

MIDAS @ ECML PKDD 2019 - The 4th Workshop on MIning DAta for financial applicationS

Würzburg (Germany), 16th September 2019

European Commission

Introduction

Recent works on the application of sentiment analysis suffer from:

- **Limited scope** of historical financial news
- Unavailability of **benchmarks** (especially long term)
- Handling of **short texts** only (usually twitter or news headlines)
- **Basic**, Natural Language Processing (NLP) **techniques** employed

Goal:

(I) considering **longer time periods**

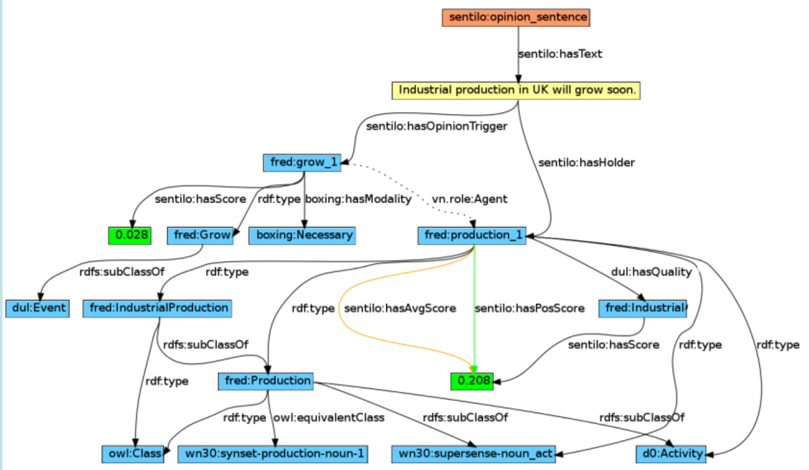
(II) analysing **entire articles**

(II) using **more sophisticated NLP** techniques

Fine-grained, aspect-based sentiment analysis

In particular we use*:

- **Fine-grained** polarity detection
- **Unsupervised** approach based on external lexical resources (sentiment dictionaries)
- **Aspect-based** sentiment analysis



- Selection of economic synonyms of an economic concept with SPARQL queries on the **World Bank Group (WBG) Ontology**
  - Classification schema of economic concepts to describe and link language and terminology used within the World Bank and beyond
    - *broader*, *narrower*, *related* relations across subject areas

* See: *D. Reforgiato Recupero, V. Presutti, S. Consoli, A. Gangemi, A.G. Nuzzolese: Sentilo: Frame-Based Sentiment Analysis. Cognitive Computation 7(2): 211-225 (2015). Demo at:* http://wit.istc.cnr.it/stlab-tools/sentilo/service

European Commission

Dow Jones DNA: Data, News and Analytics Platform

The dataset was obtained from Dow Jones and consists of several million economic and financial articles, full-text, **commercial**, since the ’94

- Considered countries: UK, US, IE, ES, IT, FR, NL, BE, DE
- Time span of **25 years**: from 01/01/1994 to 01/01/2019
- Performed a **selection** of the largest and most popular **domain outlets**

  E.g. for US: *New York Times*, *Wall Street Journal*, *Washington Post*
- Filtered subjects:

  Economic News (ECAT)

  Commodity / Financial Market News (MCAT)

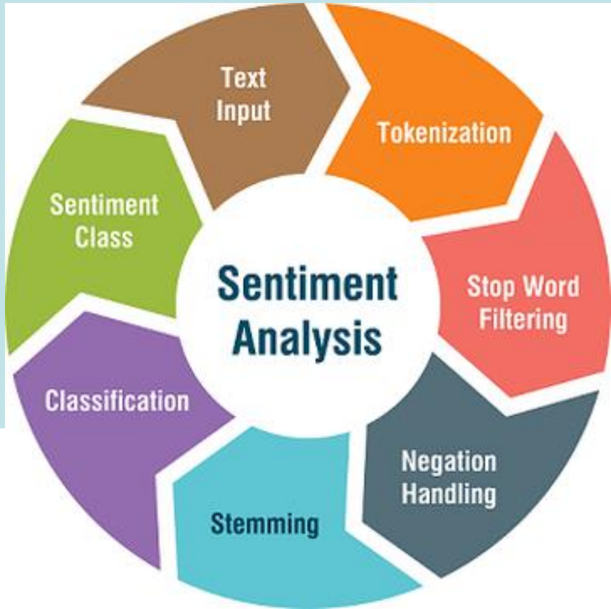
Information Extraction (IE)

- Natural Language Processing (**NLP**) pipeline to extract structured information from news that relates to search concepts of interest
- **Rule-based** IE approach based on the linguistic features of the Python library **spaCy**: Industrial-Strength Natural Language Processing
  - Looping over the POS tree stopping when our search concept, or one of its synomyms, is found
  - Navigating over the neighbouring tokens leveraging on rules based on the dependency parsing

    → Chunks of terms are constructed

For example: *…manufacturing stumbled deeper into recession...*

Manufacturing → [*stumble*, *recession*]



Rule-based approach for term chunks

- for xin in ll or xin in rr:
  - if (xin.dep_ == "**amod**" and ((xin.pos_ == "**ADJ**" and (xin.tag_ == "**JJR**" or xin.tag_ == "**JJS**" or xin.tag_ == "**JJ**")) or (xin.pos_ == "**VERB**")) …
- if (NOUN.head.pos_ == "**VERB**")…
- if (NOUN.head.pos_ == "**VERB**") and …
  - if (xin.dep_ == "**advmod**") and (xin.pos_ == "**ADV**" and (xin.tag_ == "**RBS**" or xin.tag_ == "**RBR**")…
- if (NOUN.head.pos_ == "**VERB**") and …
  - if (xin.dep_ == "**acomp**" or xin.dep_ == "**oprd**") and (xin.pos_ == "**ADJ**" and (xin.tag_ == "**JJR**" or xin.tag_ == "**JJS**" or xin.tag_ == "**JJ**")…
- if (NOUN.head.pos_ == "**VERB**") and …
  - if (xin.dep_ == "**dobj**" or xin.dep_ == "**attr**") and xin.pos_ == "**NOUN**": …
- if (NOUN.head.pos_ == "**VERB**") and …
  - if (xin.dep_ == "**xcomp**" or xin.dep_ == "**advcl**") and xin.pos_ == "**VERB**":
- if (xin.dep_ == "**acl**") and (xin.pos_ == "**VERB**")…
- …
- …
- …

Heuristics and sentiment polarity propagation

Heuristics

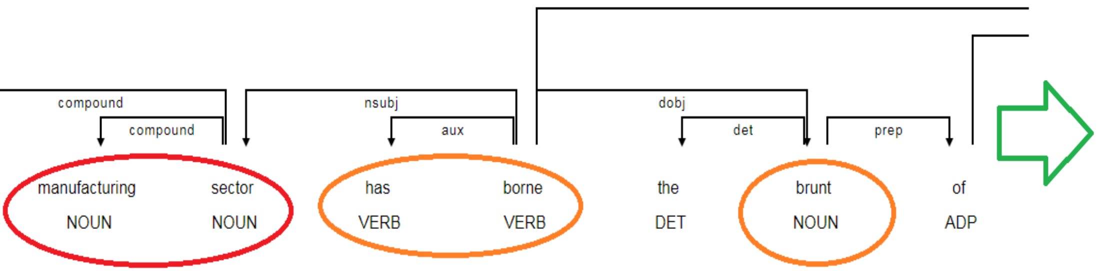
- Discovery of most frequent location
- Tense detection
- Lexical resources for sentiment scoring
  - Sentiment polarity of a term is taken from a custom economics vocabuary we are building, or from *SentiWordNet*
  - Sign consistency check with *Loughran-McDonald dictionary*

Sentiment polarity propagation

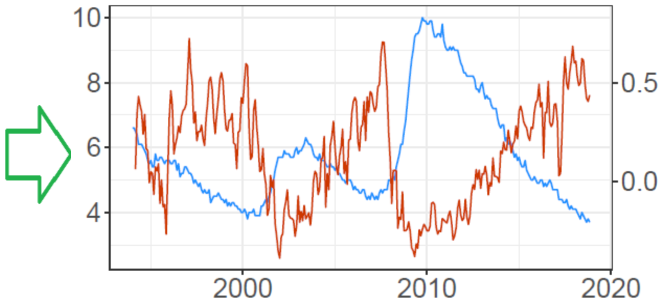
- Sentiment scores of the terms contained in a chunk are propagated to the root search concept, providing a final polarity score for it

Sentiment polarity propagation: an example

- *British manufacturing sector has borne the brunt of the global economic slowdown over the past few months and…*
  - Detected “**manufacturing sector**” by looping on part-of-speech tags
  - It is attached to a VERB: “**to bear**”
  - VERB linked to a DOBJ (direct object) which is a NOUN: “**brunt**”
  - Polarity propagation: “**brunt**” → “**to bear**” → “**manufacturing sector**”
  - Final aspect-based polarity of “**manufacturing sector**” : **-0.52**



“brunt” → “to bear” → **“manufacturing sector” → -0.52**



Preliminary analysis - US GDP

Investigate forecasting power of news to nowcast the GDP of the United States

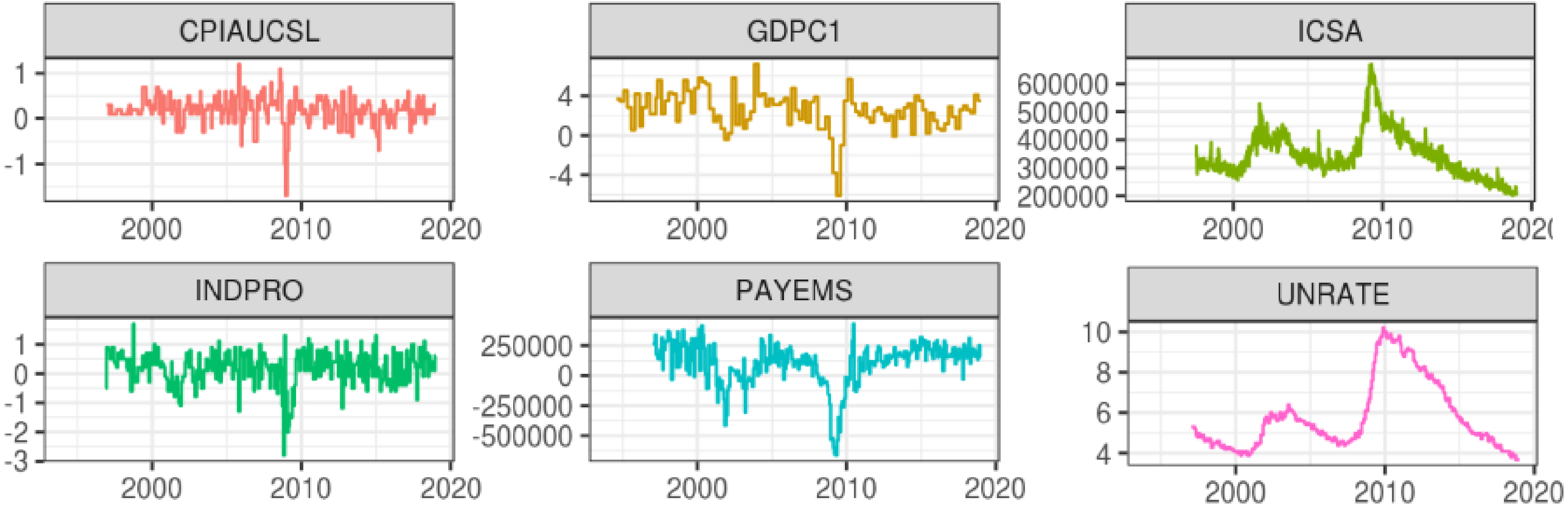
Models:

- *AR* → $y_{r,0}=\beta_0+\beta_1 y_{r,d}+\epsilon_r$
- *ARX* → $y_{r,0}=\beta_0+\beta_1 y_{r,d}+\beta_1 x_{r,d}+\epsilon_r$
- *ARN* → $y_{r,0}=\beta_0+\beta_1 y_{r,d}+\beta_1 N_{r,d}+\epsilon_r$
- *ARXN* → $y_{r,0}=\beta_0+\beta_1 y_{r,d}+\beta_1 x_{r,d}+\beta_1 N_{r,d}+\epsilon_r$
- *SS* → subset selections of the most important variables (all news indicators provided)
- *LASSO* → lasso selection and re-estimation of the linear model with the selected predictors

News sentiment for *Industrial Production*, *Monetary Policy*, *Unemployment*, *Inflation*

Different verbal forms: *past*, *present*, *future*, *NaN*

Variables considered and publication lags



LEGEND

ICSA:
Unemployment Insurance Weekly Claims

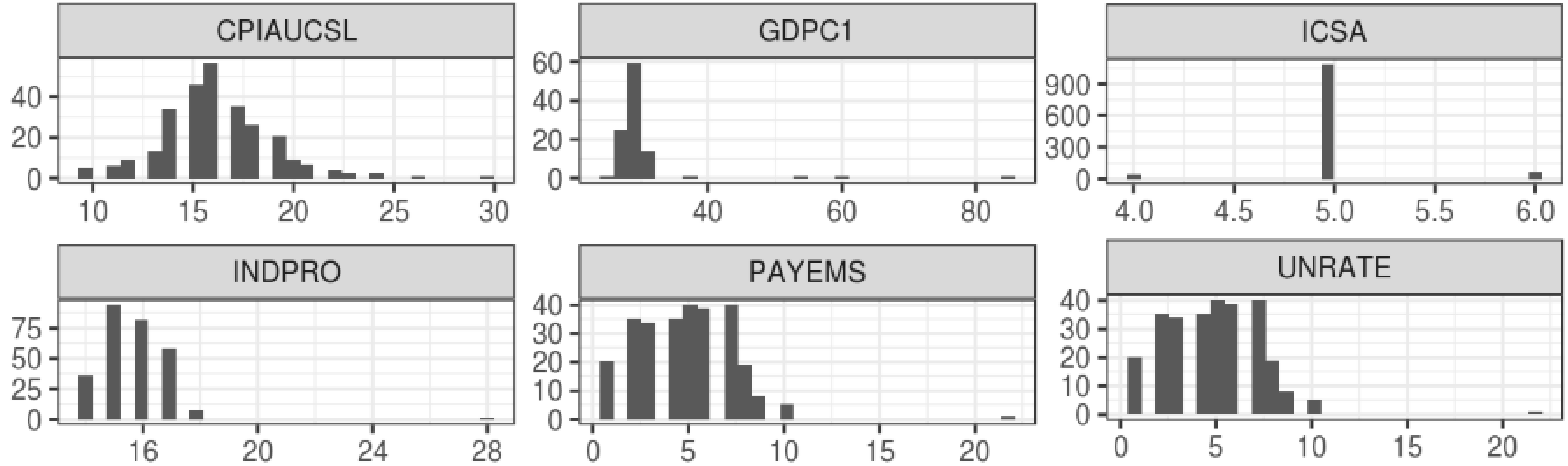
PAYEMS:
All Employees: Total Nonfarm Payrolls

CPIAUCSL:
Consumer Price Index for All Urban Consumers

UNRATE:
Unemployment Rate

GDPC1:
Real Gross Domestic Product

INDPRO:
Industrial Production Index



In-sample analysis - R^2

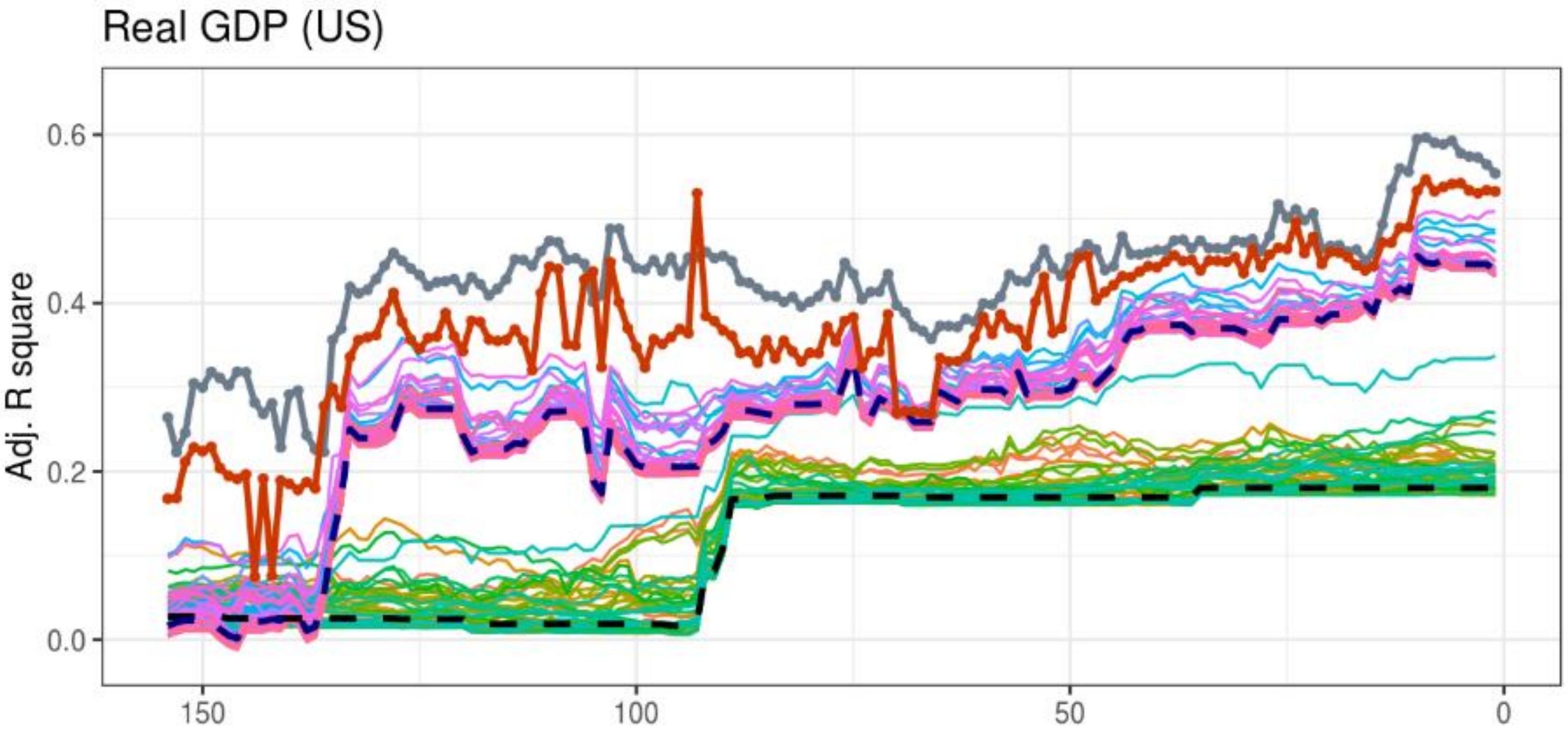
Real GDP annualized QoQ

Evaluation: R^2 of the models forecasting the release value based on the information available *d* days before

Black lines: *AR* (bottom) and *ARX* (top)

Colored lines: *ARN* and *ARXN*

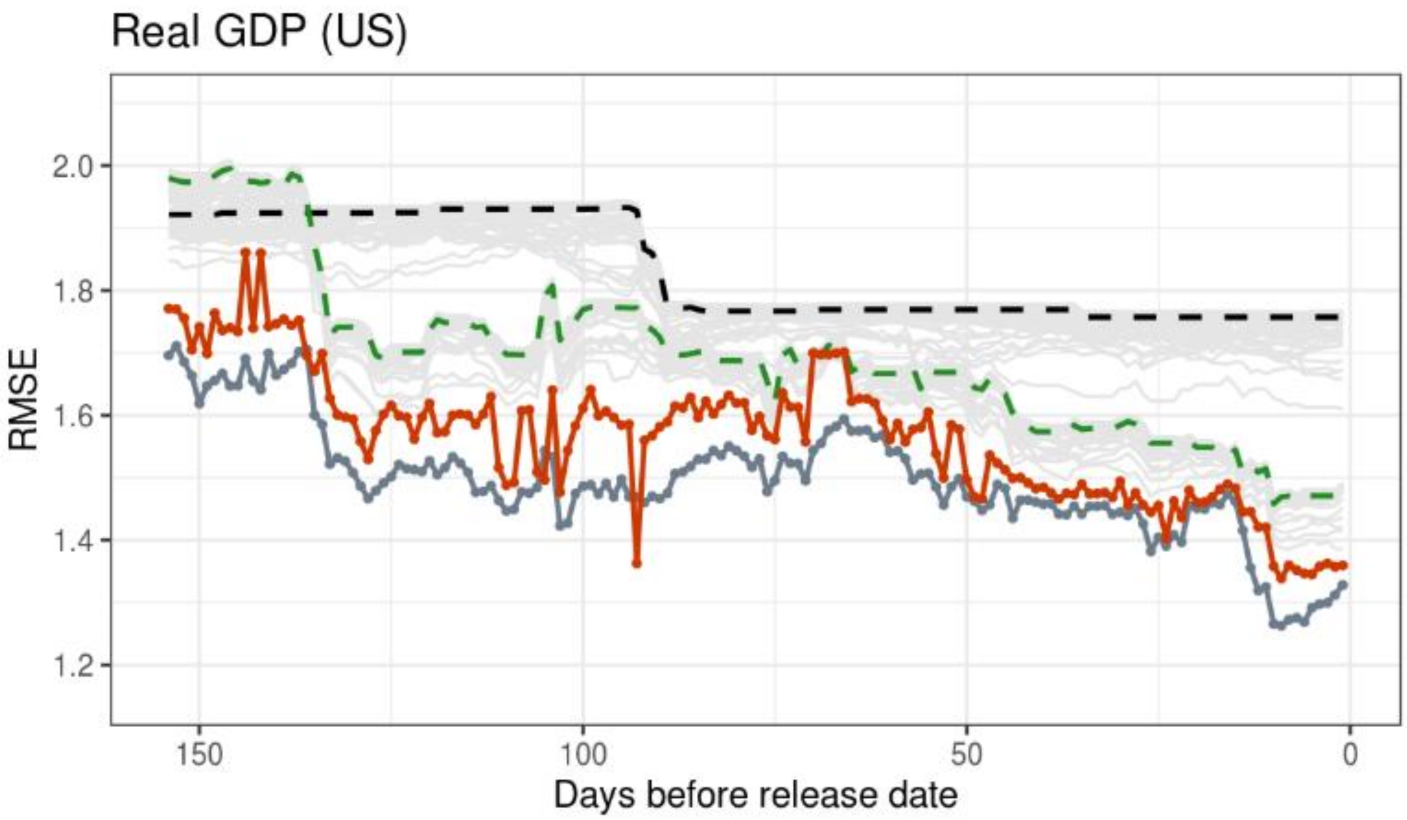
Dotted lines: *SS* (gray) and *LASSO* (orange)



In-sample analysis - *RMSE*

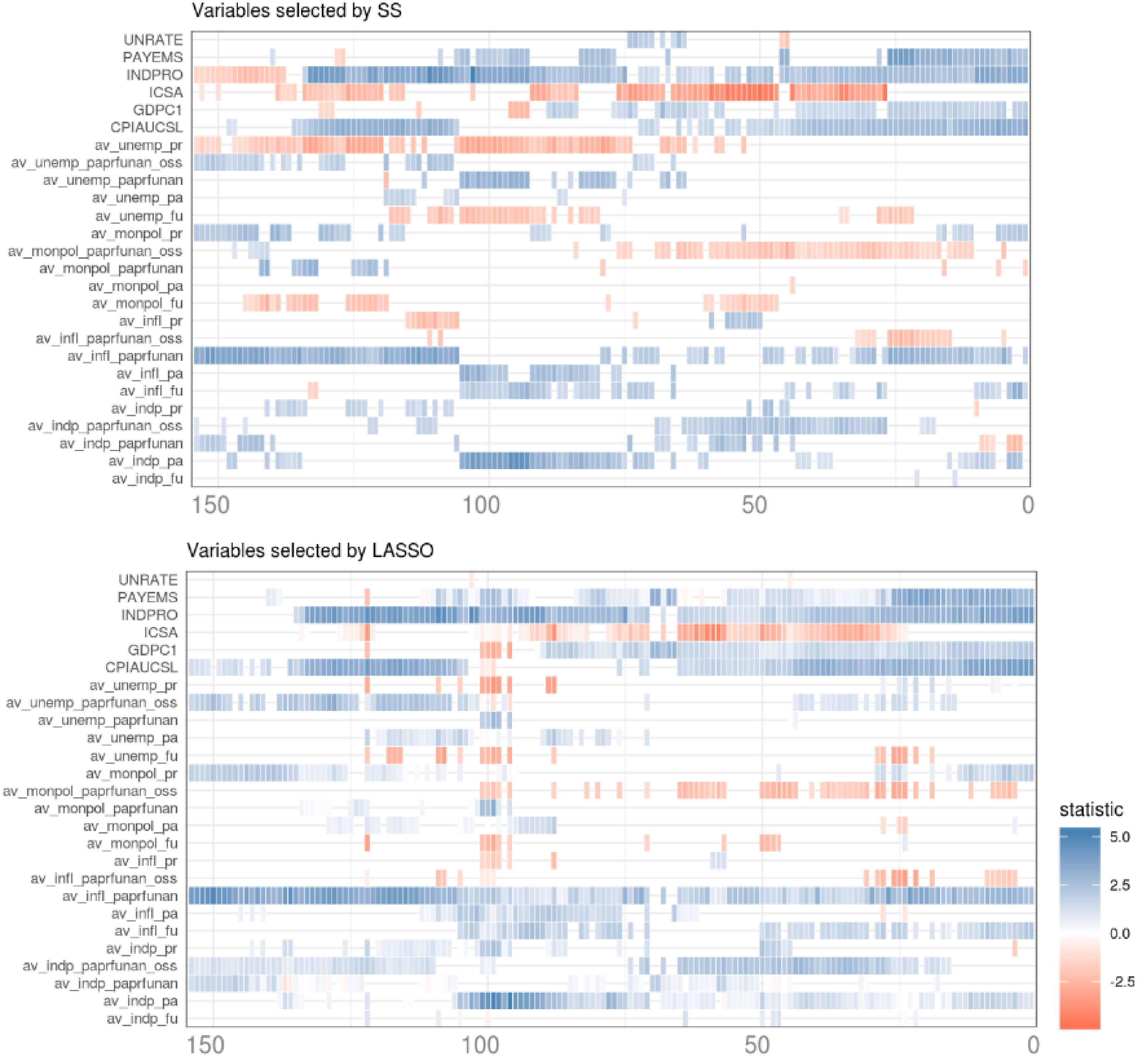
Real GDP annualized QoQ

Evaluation: *RMSE* of models forecasting the release value based on the information available *d* days before
Black lines: *AR* (top) and *ARX* (bottom)
Colored lines: *ARN* and *ARXN*
Dotted lines: *SS* (gray) and *LASSO* (orange)



In-sample analysis - *Variables selected*

Real GDP annualized QoQ



Conclusions and on-going work

- Sentiment signals extracted from economic news have significant effect on forecasting of US GDP
- No particular effects on discrimination of verbal forms (further investigation needed)
- On-going:

1. Construction of dictionary of sentiment scores for the economic/financial sector (i.e. a fine-grained extension of the Loughran-McDonald dictionary) using Mechanical Turk
2. Increasing number of news articles to be able performing out-of-sample analyses
3. Extend analysis to other countries (Europe)
4. Forecast other economic indicators: Industrial Production, Inflation, Unemployment, …
5. Supervised approach via GloVe word embedding and Machine Learning

Any questions?

You can find me sergio.consoli@ec.europa.eu

European Commission