News and banks’ equities: do words have predictive power?

Valerio Astuti, Giuseppe Bruno, Sabina Marchetti, Juri Marcucci [1]

Banca d’Italia

October 21, 2019

[1] Disclaimer: The views expressed in this presentation are the authors’ only and do not necessarily reflect those of the Bank of Italy.

Motivations and main steps

- Can we extract useful quantitative information from narrative content in newspaper?
- Are news a predictive factor in banks' equities trends?
- Is there any advantage in putting together text and classical banking balance sheet indicators?

Our steps on the main road

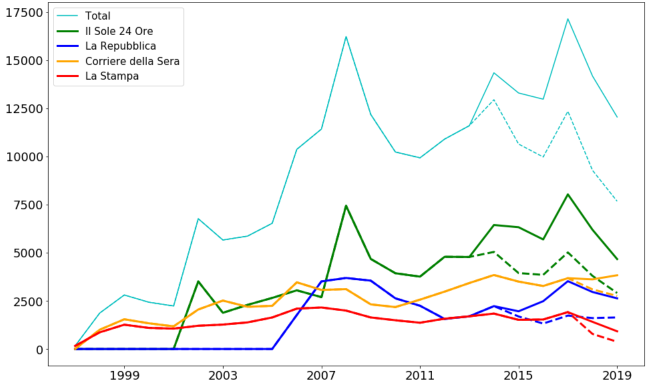
- Starting point: building an archive of articles speaking of main Italian banks
- Preprocessing of articles, topic and sentiment analysis
- Application to predictive model for banks trade volumes
- Memory-intensive tasks: Python and PySpark
- Work in progress!

Database

- From Dow Jones Factiva news aggregator we selected articles regarding 100 most important Italian banks
- Available period going from September 1996 to May 2019 (article number not uniformly distributed over time)
- Sources: “Il Sole 24 Ore”, “La Stampa”, “La Repubblica”, “Corriere della Sera” plus online edition

Database

Number of articles per source



Polishing our Corpus

- 217k articles, 100M words, 0.33M of unique tokens (Zipf’s law)
- Case normalization, tokenization, stop-words removal, stemming
- Cut-off on minimum number of appearances of a term

Latent Dirichlet Allocation (LDA)

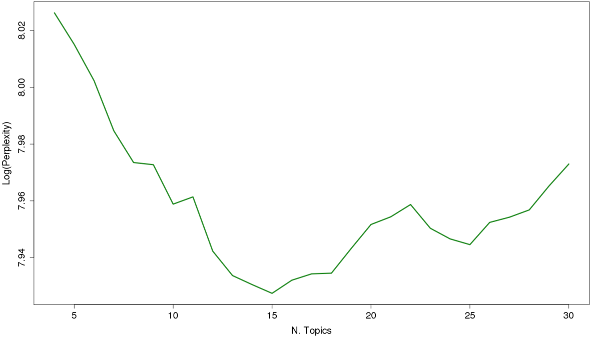
- Unsupervised, hierarchical probabilistic model to decompose a document in its most salient topics (probability distribution over words)
- Full sample (synchronic) and rolling subsample (diachronic): one defined over whole period, the other limited to three-year spans, rolling yearly. The number of topics is chosen to minimize perplexity
- Rolling sample was used to sidestep two possible problems: look-ahead bias and coarsening of topics over a 20-year span

LDA: results

- Full sample:
  minimum perplexity = 7.93,
  number of topics = 15

- Rolling sample:
  average perplexity = 7.83,
  average number of topics = 8.75

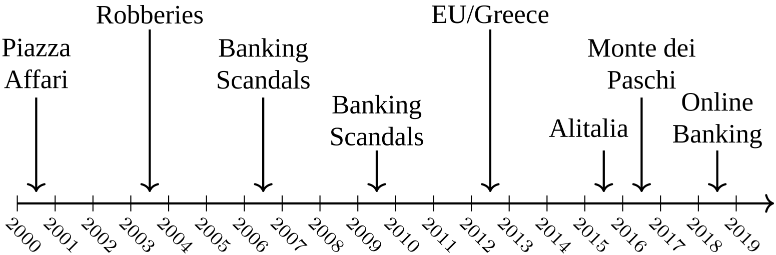
Perplexity as a function of the number of topics



LDA: main topics (full sample)

- Economy-Politics
- Investigations
- Industry
- Stock Exchange
- Local Activities
- English articles
- News Reports
- Financial Activities
- Italian Groups
- Public
- Balance/Capital
- Growth and Taxes
- Investments
- Stock Market Trends
- Boards

LDA: main topics (rolling subsample)



Piazza Affari

Robberies

Banking Scandals

Banking Scandals

EU/Greece

Alitalia

Monte dei Paschi

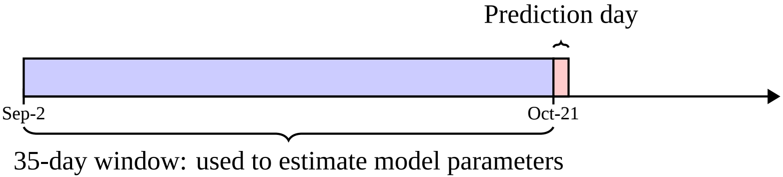
Online Banking

Model

- Topics predictive power tested on stock market indices of 4 relevant Italian banks and the Italian stock index FTSE MIB
- Studied returns, volatilities, volumes; volumes are the most reactive variables to news, and the ones topics forecast best
- Applied topic distributions with both static and rolling samples, with different results
- Our model is a LASSO with an adaptive number of topics k_t updated daily and the possibility to keep up to three lagged variables
- The benchmark model is an $AR\left(p_t\right)$ with p_t selected to minimize the BIC

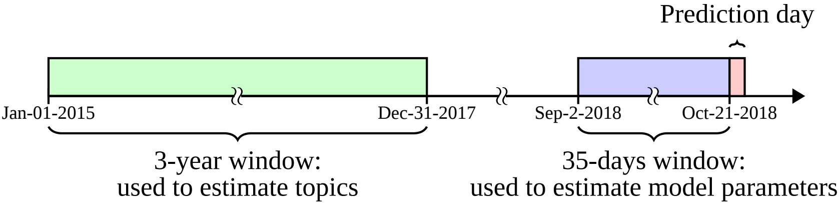
Topic distribution (full sample)

- In the static setting the predictive variables are the topics (defined over the whole available period) weighted with the daily sentiment
- The number of variables used as predictors in a given day is selected by the LASSO methodology over the previous 35-day period
- Possibility of look-ahead bias using topic estimated with future articles but on the other hand topics much coarser given the definition on a longer timespan



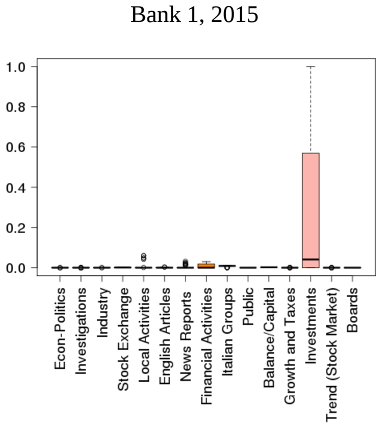
Topic distribution (rolling subsample)

- In the topic model with rolling sample the predictive variables are estimated over the 3 calendar years preceding the prediction day (weighted with the daily sentiment).
  Example: to make prediction for October 21, 2018 we would use topics estimated between January 2015 and December 2017
- Having defined the topics, the regression coefficients are estimated in the 35 days before the forecast. There the LASSO automatically selects the significant (weighted) topics



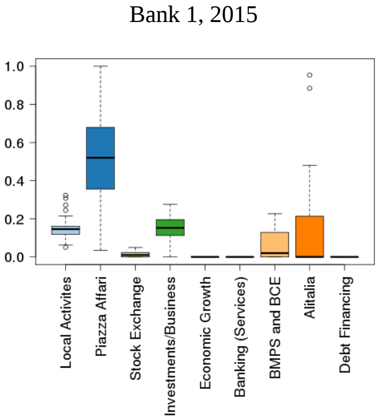
Predictive model: coefficients

Full sample



Bank 1, 2015

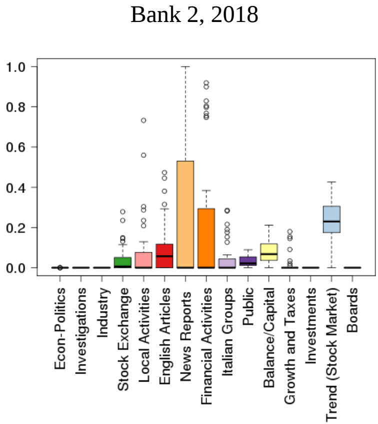
Rolling subsample



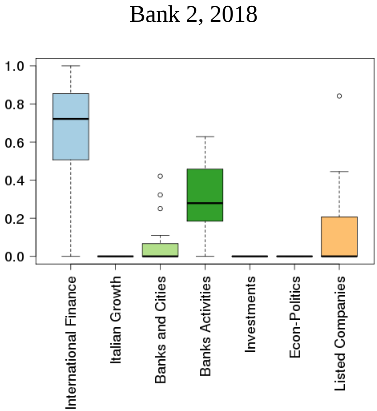
Bank 1, 2015

Predictive model: coefficients

Full sample



Rolling subsample

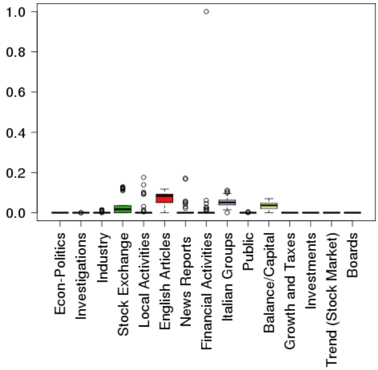


Predictive model: coefficients

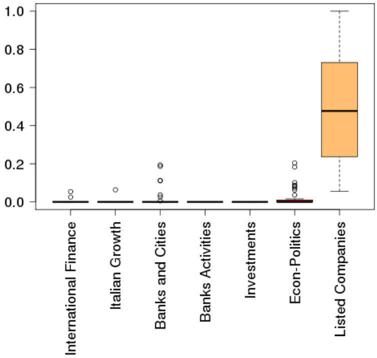
Full sample

Rolling subsample

FTSE, 2018



FTSE, 2018

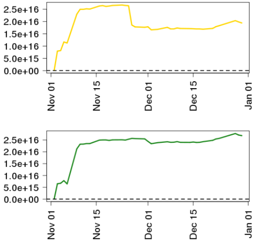


Predictive model: performance

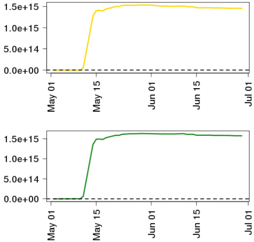
The performance of our models is evaluated through the difference between the cumulated sum of squared errors of the benchmark and our models

The upper panel is relative to the model with topics defined on the full sample, the lower to the one with topics defined on the rolling subsample

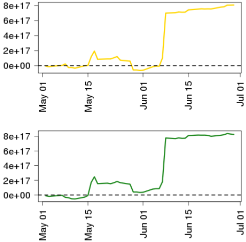
Bank 1, 2015



Bank 2, 2018



FTSE, 2018



Conclusions

- Topic analysis effectively captures relevant information content of newspaper articles
- Topics can be used as predictor variables for banks' equities
- Using the topics to make predictions our models perform on average better than the autoregressive benchmark

Thank You for Your Attention