Deriving Indicators from a Large Corpus of Italian Documents
&
News and Banks' Equities: Do Words Have Predictive Power?

Discussed by Federico Maria Ferrara

European Central Bank

"Big Data and Machine Learning Applications for Central Banks" Workshop

Bank of Italy - 21 October 2019

Relevance of the Two Projects

1. The projects speak to an increasing literature that analyzes news text for stock price prediction and nowcasting macroeconomic variables (for a review, Gentzkow et al. 2019).
   - Most of the literature relies on dictonary-methods and word counts (e.g., Tetlock 2007; Loughran and McDonald 2011).
   - There is room to exploit machine learning based techniques of text analysis to improve economic forecasts.

Relevance of the Two Projects

1. The projects speak to an increasing literature that analyzes news text for stock price prediction and nowcasting macroeconomic variables (for a review, Gentzkow et al. 2019).
    - Most of the literature relies on dictonary-methods and word counts (e.g., Tetlock 2007; Loughran and McDonald 2011).
    - There is room to exploit machine learning based techniques of text analysis to improve economic forecasts.

2. "*The field of economics should be expanded to include serious quantitative study of changing popular narratives*" (Shiller 2017).
    - Topic modeling is a promising direction for the rigorous assessment of narratives' impact on economics fluctuations.

Main Results

1. Based on large corpora of Italian news, the projects derive time series of latent topics and dictionary-based sentiment scores.
    - LDA appears to effectively capture relevant information content.

Main Results

1. Based on large corpora of Italian news, the projects derive time series of latent topics and dictionary-based sentiment scores.
    - LDA appears to effectively capture relevant information content.

2. (Some) topic-augmented models show enhanced forecasting performance vis-à-vis naïve models.

Strengths

1. **Data**

   $\Rightarrow$ The projects make use of novel and rich datasets.

   $\Rightarrow$ The use of Factiva DNA is the state of the art for text data retrieval.

Strengths

1. **Data**
   - ⇒ The projects make use of novel and rich datasets.
   - ⇒ The use of Factiva DNA is the state of the art for text data retrieval.

2. **Transparency**
   - ⇒ The projects describe in a well-structured and replicable manner their pipelines from query construction to model specification.
   - ⇒ Efforts are put to minimize the degree of arbitrariness of important reseach design decisions (e.g., number of topics in LDA model).

Strengths

1. **Data**
   - ⇒ The projects make use of novel and rich datasets.
   - ⇒ The use of Factiva DNA is the state of the art for text data retrieval.
2. **Transparency**
   - ⇒ The projects describe in a well-structured and replicable manner their pipelines from query construction to model specification.
   - ⇒ Efforts are put to minimize the degree of arbitrariness of important reseach design decisions (e.g., number of topics in LDA model).
3. **Visualisation**
   - ⇒ LDA results are visualised in a clear and informative manner.

Room for Improvement: Ex Ante Decisions

1. **Where is theory?**
   - ⇒ The reason why certain topics have more predictive power than others is a black box.
   - ⇒ Is it possible to have a more "theory-informed" approach and hypothesis-testing?

Room for Improvement: Ex Ante Decisions

1. **Where is theory?**
   - ⇒ The reason why certain topics have more predictive power than others is a black box.
   - ⇒ Is it possible to have a more "theory-informed" approach and hypothesis-testing?

2. **Text preprocessing**
   - ⇒ Denny and Spirling (2018) show that preprocessing decisions have profound effects on the results of unsupervised learning models.
   - ⇒ Is it possible to minimize and standardize the amount of preprocessing choices (e.g., bi-gram inclusion, document-term matrix trimming)?

Room for Improvement: Ex Ante Decisions

1. **Where is theory?**
   - ⇒ The reason why certain topics have more predictive power than others is a black box.
   - ⇒ Is it possible to have a more "theory-informed" approach and hypothesis-testing?

2. **Text preprocessing**
   - ⇒ Denny and Spirling (2018) show that preprocessing decisions have profound effects on the results of unsupervised learning models.
   - ⇒ Is it possible to minimize and standardize the amount of preprocessing choices (e.g., bi-gram inclusion, document-term matrix trimming)?

3. **Unsentimental sentiment**
   - ⇒ "*[Off-the-shelf] dictionaries are able to produce measures that are claimed to be about tone or emotion, but the actual properties of these measures – and how they relate to the concepts they are attempting to measure – are essentially a mystery*" (Grimmer and Stewart 2013).

Room for Improvement: Model Selection

1. **Perplexed about "perplexity"**
   - ⇒ Held-out likelihood is not (or is negatively) correlated with human judgement (Chang et al. 2009).
   - ⇒ Paradox: models with better statistical fit have worse topic interpretability.
   - ⇒ Is this an ultimately informative metric to evaluate LDA performance?

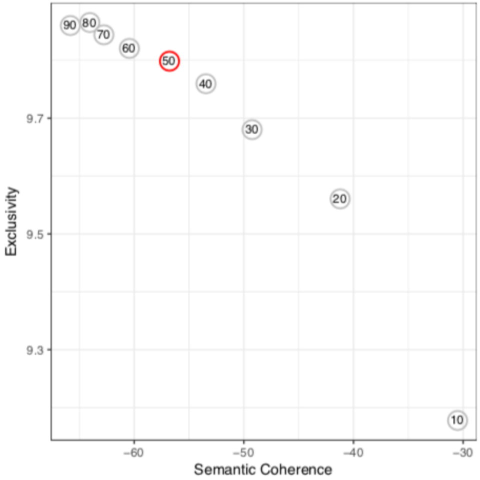
Room for Improvement: Model Selection

1. **Perplexed about "perplexity"**
   - ⇒ Held-out likelihood is not (or is negatively) correlated with human judgement (Chang et al. 2009).
   - ⇒ Paradox: models with better statistical fit have worse topic interpretability.
   - ⇒ Is this an ultimately informative metric to evaluate LDA performance?

2. **Coherence-exclusivity trade-off**
   - ⇒ Roberts et al. (2014) propose to measure topic quality through a combination of semantic coherence and exclusivity of words to topics.
   - ⇒ FREX metric (Bischof and Airoldi 2012) is used to measure exclusivity in a way that balances word frequency.
   - ⇒ Coherence and exclusivity are inversely proportional. Worth considering this trade-off in model selection.

Coherence-Exclusivity Trade-Off: An Example

Figure A1: Exclusivity and Semantic Coherence Measures for Varying Numbers of Topics



NOTES: This figure shows exclusivity and semantic coherence scores for nine topic models estimated on the German corpus of newspaper articles. The number associated with each observation corresponds to the number of topics included for each model whose exclusivity and semantic coherence is reported.

Room for Improvement: Evaluation and Prediction

1. **Look-ahead bias**

   $\Rightarrow$ The second project tries to address the problem with a rolling sample, but this yields relatively worse predictions.

   $\Rightarrow$ Is forecasting performance driven by information that is not available during the time period being simulated?

Room for Improvement: Evaluation and Prediction

1. **Look-ahead bias**

   ⇒ The second project tries to address the problem with a rolling sample, but this yields relatively worse predictions.

   ⇒ Is forecasting performance driven by information that is not available during the time period being simulated?

2. **Time period**

   ⇒ In the first project, the sample of articles goes from 1996 to 2019.

   ⇒ Why only the 2007-2019 sample is considered for inflation forecasts?

Room for Improvement: Evaluation and Prediction

1. **Look-ahead bias**
   - ⇒ The second project tries to address the problem with a rolling sample, but this yields relatively worse predictions.
   - ⇒ Is forecasting performance driven by information that is not available during the time period being simulated?

2. **Time period**
   - ⇒ In the first project, the sample of articles goes from 1996 to 2019.
   - ⇒ Why only the 2007-2019 sample is considered for inflation forecasts?

3. **Forecasting model**
   - ⇒ Similar forecasting exercises with text-based time series typically rely on VAR frameworks (e.g., Tetlock 2007).
   - ⇒ Is AR(1) too much of a "naïve" model?

Methodological Suggestions for Future Research

- **Dynamic topic models** (Blei and Lafferty 2006)
  - ⇒ They can be used to analyze the over time evolution of topics.
  - ⇒ Useful for long time frame, as words are more likely to change.
  - ⇒ *DtmModel* from *gensim* Python library.

Methodological Suggestions for Future Research

- **Dynamic topic models** (Blei and Lafferty 2006)
  - $\Rightarrow$ They can be used to analyze the over time evolution of topics.
  - $\Rightarrow$ Useful for long time frame, as words are more likely to change.
  - $\Rightarrow$ *DtmModel* from *gensim* Python library.

- **Structural topic models** (Roberts et al. 2016)
  - $\Rightarrow$ They include document-level covariate information, which can improve topic inference and qualitative interpretability.
  - $\Rightarrow$ Example of document-level covariate: media outlet (e.g., Corriere, etc.)
  - $\Rightarrow$ *stm* package in R.

Methodological Suggestions for Future Research

- **Dynamic topic models** (Blei and Lafferty 2006)
  - ⇒ They can be used to analyze the over time evolution of topics.
  - ⇒ Useful for long time frame, as words are more likely to change.
  - ⇒ *DtmModel* from *gensim* Python library.

- **Structural topic models** (Roberts et al. 2016)
  - ⇒ They include document-level covariate information, which can improve topic inference and qualitative interpretability.
  - ⇒ Example of document-level covariate: media outlet (e.g., Corriere, etc.)
  - ⇒ *stm* package in R.

- **Supervised learning**
  - ⇒ It outperforms dictionaries in sentiment analysis (Barberá et al. 2016).
  - ⇒ Labelling takes time, but it makes validation easier.
  - ⇒ *scikit-learn* library in Python and *e1071* package in R.

Thank you for your attention!

References I

Barberá, P., A. Boydstun, S. Linnand, J. Nagler, and R. McMahon (2016). Methodological Challenges in Estimating Tone: Application to News Coverage of the U.S. Economy. *Paper presented at the 2016 Annual meeting of the American Political Science Association.*.

Bischof, J. M. and E. M. Airoldi (2012). Summarizing Topical Content with Word Frequency and Exclusivity. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 9–16.

Blei, D. M. and J. D. Lafferty (2006). Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, pp. 288–296.

Denny, M. J. and A. Spirling (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis 26*(2), 168–189.

References II

Gentzkow, M., B. T. Kelly, and M. Taddy (2019). Text as Data. *Journal of Economic Literature 57*(3), 535–574.

Grimmer, J. and B. M. Stewart (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*(3), 267–297.

Loughran, T. and B. McDonald (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance 66*(1), 35–65.

Roberts, M. E., B. M. Stewart, and D. Tingley (2016). Stm: R Package for Structural Topic Models. *Journal of Statistical Software*.

Shiller, R. J. (2017). Narrative Economics. *NBER Working Paper No. 23075*.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Marketq. *The Journal of Finance 62*(3), 1139–1168.