NBER - Big Data and High-Performance Computing for Financial Economics

Language Modeling in Financial Markets

April 4, 2019

Bryan Routledge

Associate Professor of Finance
Tepper School of Business
Carnegie Mellon University
routledge@cmu.edu
http://sulawesi.tepper.cmu.edu

Vitaly Meursault

Doctoral Student
Tepper School of Business
Carnegie Mellon University
meursault@cmu.edu
https://sites.google.com/site/vitalymeursault

Proposal

We are happy to submit our proposal for your NBER call on "Big Data and High-Performance Computing for Financial Economics." While we think it fits in the broad call for proposals, it specifically fits in "Applications of machine learning and/or artificial intelligence in finance."

Language Modeling in Financial Markets

Measured by volume, most of the disclosure made by firms is text and is not numerical. Think of the narrative in the MD&A portion of a 10K document versus the quantitative earnings-per-share, return, or price. Similarly, news, blogs, commentaries, social media posts are all dominated by text, language, and discussion, but not numbers. Firms, analysts, media, and people spend a lot of time and effort creating text alongside numerical data. Of course, the volume is likely not a good measure of information content. It is an empirical question as to whether there is any information, any new information, or any information beyond that in contained in Compustat/CRSP numerical information. Addressing this question requires a statistical model of language.

There has been some success in using text as data in financial economics (Tetlock (2007), Kogan, Levin, Routledge, Sagi, and Smith (2009), Hoberg and Phillips (2016), Hoberg and Lewis (2017), Routledge, Sacchetto, and Smith (2018), Chebonenko, Gu, and Muravyev (2018)). Interestingly, the last few years has seen rapid progress in machine learning particularly in the area of Natural Language Processing. The widespread use of Siri and Alexa and the 2018 Turing Award to Yann LeCun, Geoffrey Hinton, and Yoshua Bengio for pioneering AI are both reflections of this progress. We propose to investigate and compare an array of recent natural language processing models to quantify the the information content of financial language.

To start, we use a straightforward prediction task to quantify the information content of text. On the left-hand side, we look at a variety of measures of return volatility. On the right-hand side we use the

numerical firm data and/or the text. Forecasting returns – that is, understanding the cross-section of expected returns – is difficult (Roll (1988)). Forecasting volatility is easier and provides a useful context to define the information content in language. In particular, we can see what insight the text can provide as to what the market is reacting to. We have a variety of datasets on hand and on deck (firm 10K, 10Q, earnings announcement 8K, News media text). With progress on the volatility prediction task, we aim to look at applying similar models to the cross-section of expected returns.

Specifically, we plan to use a variety of natural language processing statistical models that fall in two broad categories. The first category reduces the complexity of language by treating words (phrases) irrespective of their order – treating the text as a "bag-of-words." We look at Elastic Net text regression (Zou and Hastie (2005)) and XGboost trees (Chen and Guestrin (2016)), and Latent Dirichlet Allocation topic models (Blei, Ng, and Jordan (2003), and Roberts, Stewart, Tingley, and Airoldi (2013)). All these bag-of-words models are "linear" which allows easy inspection weights on individual words or phrases. The second class of models we estimate uses the "deep learning" architecture that underlies the recent advances in AI. Here we look at Convolutional Neural Networks (Collobert et. al. (2011)), Long Short-Term Memory models (Hochreiter and Schmidhuber (1997)), and attention-based models (Vaswani et. al. (2017)). Since these models use the words-in-order to capture the richness of the language, the models are not-linear and are harder to interpret. Here, word context matters so it is harder to see how a specific word/phrase is weighted. While it takes some work to understand the model (see Sundararajan, Taly, and Yan (2017)), the examination of words, phrases or concepts based on context or their location in the document can yield interesting insights (see Meursault (2019)).

It is not obvious that the AI revolution will translate directly to financial economics. Financial data poses unusual challenges for machine learning statistical models of text. First, disclosure and news documents are long (longer than, say, a Yelp review (Chahuneau, Gimpel, Routledge, Scherlis, and Smith (2012)). This makes deep learning models computationally harder to estimate. Second, financial data is notoriously noisy. Machine learning techniques are not out-of-the-box well suited to settings where the anticipated R2 is so low. The high dimensionality of text (possible words, possible sentences), makes it harder to implement the standard variance deducing "sorting" techniques used in cross-sectional asset pricing. Finally, financial disclosure setting is strategic and may not be well captured by a reduced-form empirical framework. Structural models of language that incorporate the writers' objectives would be interesting to implement on the financial economic data we are using here (see Sim, Routledge, and Smith (2015a), Sim, Routledge, and Smith (2015b), and Sim, Routledge, and Smith (2016)).

Finally, the project will be computation-heavy. The data we are using is relatively large, 50,000 to 200,000 observations depending on the task. Documents tend to be relatively long, several thousand tens of thousands of words. Estimating deep learning models comes with a heavy computational burden. A large part of the success of recent AI models has been due to improvements in estimation procedures driven by stochastic gradient descent methods. However, they still typically require (or benefit from) a specialized GPU computer setting like, for example, the Pittsburgh Supercomputing Center we have used in prior research.

References

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Chahuneau, V., K. Gimpel, B. R. Routledge, L. Scherlis, and N. A. Smith (2012): "Word Salad:

Relating Food Prices and Descriptions," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1357–1367, Jeju Island, Korea. Association for Computational Linguistics.

Chebonenko, T., L. Gu, and D. Muravyev (2018): "Text sentiment's ability to capture information: evidence from earnings calls," *Working paper*.

Chen, T., and C. Guestrin (2016): "XGBoost: a scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp. 785–794.

Collobert, R., J. Weston, L. Bottou, K. Kavukcuoglu, and P. Kuksa (2011): "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, 12, 2493–2537.

Hoberg, G., and C. Lewis (2017): "Do fraudulent firms produce abnormal disclosure?," *Journal of Corporate Finance*, 43, 58–85.

Hoberg, G., and G. Phillips (2016): "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, 124(5), 1423–1465.

Hochreiter, S., and J. Schmidhuber (1997): "Long short-term memory," *Neural Computation*, 9(8), 1735–1780.

Kogan, S., D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith (2009): "Predicting Risk from Financial Reports with Regression," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280, Boulder, Colorado. Association for Computational Linguistics.

Meursault, V. (2019): "The language of earnings announcements," *Working paper*.

Roberts, M. E., B. M. Stewart, D. Tingley, and E. M. Airoldi (2013): "The structural topic model and applied social science," *NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation.*

Roll, R. (1988): "R2," *The Journal of Finance*, 43(3), 541–566.

Routledge, B. R., S. Sacchetto, and N. A. Smith (2018): "Predicting merger targets and acquirers from text," *Working paper*.

Sim, Y., B. Routledge, and N. A. Smith (2015a): "A Utility Model of Authors in the Scientific Community," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1510–1519, Lisbon, Portugal. Association for Computational Linguistics.

Sim, Y., B. Routledge, and N. A. Smith (2015b): "The utility of text: The case of Amicus briefs and the Supreme Court," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2015)*, Austin, Texas. AAAI.

Sim, Y., B. Routledge, and N. A. Smith (2016): "Friends with Motives: Using Text to Infer Influence on SCOTUS," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1733, Austin, Texas. Association for Computational Linguistics.

Sundararajan, M., A. Taly, and Q. Yan (2017): "Axiomatic Attribution for Deep Networks," *ICML'17 Proceedings of the 34th International Conference on Machine Learning*, 70(3319-3328).

Tetlock, P. C. (2007): "Giving content to investor sentiment: the role of media in the stock market," *The Journal of Finance*, 62(3), 1139–1168.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017): "Attention is all you need," *Advances in Neural Information Processing Systems 30*, pp. 5998–6008.

Zou, H., and T. Hastie (2005): "Regularization and variable selection via the elastic net," *J. R. Statist. Soc. B*, 67(2), 301–320.