

A hypothesis about the performance-gain delivered by supervised topic models over unsupervised topic models for historical archives as corpus, pertaining to secular versus non-secular changes in the response variable.

Sayan Bhattacharyya
July 29, 2011

The article "[Money, Prices and Wages in the Confederacy, 1861-65](#)" by Eugene Lerner (Lerner 1955) initiated a debate in which Lerner described the increasing inflation in the Confederacy over the course of the Civil War, and analyzed its reasons, attributing it to the increase in the stock of paper money over the duration of the war. This stock kept rising (Lerner 1955, p. 20, Table 1) because of several factors: Firstly, owing to the war, factories were not only being destroyed in the South, but also, due to the disruption of supply routes, those factories that still functioned were being starved of their raw materials; as a result of both of these factors, money that otherwise would have been invested into production could not be so invested, and hence remained in circulation, increasing the circulating money stock over time. Secondly, to fund the war effort, the Confederacy also kept minting new money, which further swelled the existing stock of money in circulation. Lerner also gives an estimation of what the month-by-month consumer price index (computed overall, that is, for the entire Confederacy) had been for the duration of the Civil War, computing the equivalent of a consumer price index series based on the quotations for each of a carefully selected mix of common commodities, such that each commodity in the mix had been continuously listed in the newspapers during the Civil War's duration (Lerner 1955, p. 24, Figure 1 and Table 2). Because inflation was constant and uninterrupted, the CPI values in the series kept increasing throughout the war.

Lerner's argument, namely, that the stock of money in circulation was the primary reason for the rise in prices of commodities, conforms well with the monetarist argument that the role of money supply is a paramount causal factor in understanding economic phenomena (Brunner & Meltzer, 1993). In fact, Milton Friedman advanced the Confederate situation during the Civil War as a textbook example of how severe changes in money supply cause correspondingly serious changes in prices (Friedman 1992). Subscribing to this view entails believing that transient, day-to-day events will have a much lower impact on prices than will the stock of money in circulation. This view, however, has been challenged in the ensuing years since Lerner's paper was published: George McCandless, Jr. argues, for example, that local phenomena including war news had a significant on prices (McCandless 1996), while even more recent work has painted a more nuanced picture (Burdekin & Weidenmier 2001). Of course, it is not necessarily the case that the two explanations cannot co-exist. The more "materially" oriented explanation of price rise by way of increase in the stock

of money in circulation is a macro-level, aggregate explanation, over the range of the entire Confederacy as a whole. On the other hand, the more psychologically based explanation of price rises as influenced by individuals reacting to unpropitious local events is based on what is arguably much more of a subjective and cultural phenomenon. It is no doubt undeniable that even the money-supply based “materialist” explanation ultimately depends on the mediation of a subjective component (it is, after all, individual people who make the decision to raise or lower prices), but the subjective component in it is of lesser importance because of the fact that the money-supply is a macro-economic variable -- whereas, in that epoch before mass media in which news traveled relatively slowly, even calamitous events may have had far less impact outside of a fairly limited local context.

We are neither historians nor economists, and our interest in this matter comes from a different direction. We are interested in this particular debate among economic historians of the Civil War because it affords us an opportunity to explore a particular aspect of topic models, which are a relatively new instrument in discourse analysis [A short description of what topic models are, goes here.] In the humanities and social sciences, topic models have so far been applied to literary scholarship [Mimno ????, Riddell 2011] and by historians [Nelson 201?] to discover themes in large corpora. Like any other inductive learning technique, topic models have been shown to perform better with supervised learning as compared to unsupervised learning (Blei & McAuliffe 2007). Supervised learning, however, is more expensive than unsupervised learning, as a suitable response variable needs to be researched and the data needs to be trained with [How, exactly, does the training process work? I am unclear as to the specifics of the supervision mechanism in our topic models.] Clearly, it would be helpful to develop general rules of thumb as to situations in which putting in the extra effort to generate supervised topic models as opposed to unsupervised topic models will be worth the extra effort. However, there has not so far been much exploration of this in the research literature. In the present work, we will test an intuition about the general situation in which supervised topic models may be expected to lead to diminishing returns over unsupervised topic models.

Before we approach that question, let us briefly recapitulate how supervised learning enhances the performance of topic models over unsupervised learning. In supervised topic models, a response variable is associated with the training data, while in unsupervised topic models, there is no such response variable. In a supervised topic model, therefore, the learning that is encoded by the model encompasses the association of topics with the response variable. In other words, the topics discovered in supervised learning have a selection bias in favor of the degree of their relevance to the response variable in question. The topics discovered through supervised learning, therefore, are likely to be more homogeneous. That, of course, begs the question as to what it means for topics to be more homogeneous. It could mean one of two

possibilities: The first possibility is that topics discovered through supervised learning may be more homogeneous with respect to *each other*, i.e. the topics discovered may be more self-similar, or “consistent”. In other words, the homogeneity induced by the supervision mechanism may be a homogeneity *across* topics. Another possibility is that each of the topics discovered through supervised learning may be more homogeneous with respect to *itself*, that is, the *words* in any particular individual topic that is discovered, happen to be more self-similar -- making the topics themselves more “coherent”. In such a case, the homogeneity induced by the supervision mechanism would be *within* each topic, rather than *across* topics.

To come back to our question, we are interested in knowing when it is that supervised topic models may be expected to lead to diminishing returns over unsupervised topic models. For simplicity's sake, let us limit ourselves to situations in which the response variable (R) represents a metric – for example, the “starred” rating of a film in a film review (from one to five stars), or (as in our case), the Consumer Price Index (CPI) for a particular month. We can, then, express the following intuition:

(1) The performance-quality gain added by the supervision mechanism of a supervised topic model (as compared to an unsupervised topic model) would depend, other things being equal, on the extent to which the correlation between the response variable R and the data can be captured.

This is where data from a corpus such as an archive of newspaper articles (consisting of roughly the same number of articles per unit of time) becomes very interesting to consider as a dataset. We have two intuitions about such an archive:

(2) If the response variable R is a quantity that is important to people's lives (such as, for example, the price of needed commodities), then fluctuations in R are likely to be reflected as a corresponding fluctuation of *some* form (for our purposes, we can call it “note-taking”) in the newspaper articles – that is, the newspaper articles would, in some fashion, make note of the changes, especially at moments when the change is for the worse (since bad news is more “newsworthy” – as the saying goes, “no news is good news”) – thus, for example, an abrupt rise in the CPI is a “newsworthy” event;

(3) If it is the case, however, that the response variable R shows a long-term (secular) trend, then, however, the changes brought about by such a trend are likely to be less “newsworthy”. We can think of this as the “frog-in-boiling-water” rule: it is said that, if you keep raising the temperature of water until it reaches boiling point, a frog immersed in it would not even notice it as long as the temperature is raised gradually and at a constant rate. Similarly, if what makes news is a departure from expectation, then changes brought about by an ever-present, secular trend are not likely to be newsworthy. If that is the case, then fluctuations in R that are *brought about by a secular trend* are *less* likely to be reflected as a corresponding “note-taking” in the

newspaper articles than the fluctuations in **(1)** above.

If the intuitions **(2)** and **(3)** above are correct, then it follows that the correlation between the response variable R and the corpus will be less pronounced in the case of **(3)**, that is, when R changes in response to factors not having to do with a secular trend, than in the case of **(2)**, that is, when R changes in response to a secular trend. It then follows (from the intuition **(1)** above) that the performance-quality gain added by the supervision mechanism of a supervised topic model (as compared to an unsupervised topic model) would be less in the case of **(3)**, that is, in the case of a secular-trend, than in the case of **(2)**, that is, in the case of a non-secular trend (**Hypothesis 1**). We can think of this as our null hypothesis.

By a stroke of good luck, the available CPI data from the Civil War years affords us a quite straightforward way to test **Hypothesis 1**. There is, indeed, a pronounced secular upward trend for the CPI for the Confederacy over the entire course of the war [provide the reference here] – cost of living, averaged over the Confederacy, increased almost monotonically throughout the Civil War years. However, data is also available that shows *non-secular* variations in the CPI of individual cities within the Confederacy, during this period. Lerner provides a comparison of *Richmond's prices* to prices for *all Southern cities* [(Lerner 1955 p. 25, Fig. 3) – thus, CPI data that is immediate to the city of Richmond, the place of publication of the newspaper (the *Richmond Daily Dispatch*) whose archive for these years constitutes our corpus. This Richmond-specific CPI shows many fluctuations but no secular trend.

It follows, then, from our **Hypothesis 1** that we would expect there to be a higher performance-quality gain obtained by the use of a supervised mechanism over an unsupervised mechanism in the case of topic models using the Richmond Daily Dispatch archive for the Civil War years as the corpus and the Richmond-specific CPI as the response variable, than in the case of topic models using the same corpus but with the average CPI for all cities as the response variable.

References:

Blei, David M. & Jon D. McAuliffe. 'Supervised Topic Models.' *Neural Information Processing Systems* 21. 2007.

Brunner, Karl, and Allan H. Meltzer. *Money and the Economy: Issues in Monetary Analysis*, Cambridge: Cambridge University Press. 1993.

Burdekin, Richard C. K. & Marc D. Weidenmier. 'Inflation Is Always and Everywhere a Monetary Phenomenon: Richmond vs. Houston in 1864.' *The American Economic Review*, Vol. 91, No. 5 (Dec., 2001), pp. 1621-1630.

Friedman, Milton. *Money Mischief: Episodes in Monetary History*. San Diego, CA:

Harcourt Brace Jovanovich. 1992.

Lerner, Eugene M. 'Money, Prices, and Wages in the Confederacy, 1861-65'. *Journal of Political Economy*, Vol. 63, No. 1 (Feb., 1955), pp. 20-40.

McCandless, George T., Jr. 'Money, Expectations, and the U.S. Civil War.' *American Economic Review*, Vol. 83, No. 1 (June 1996), pp. 661-71.

Nelson, Robert K. "[Mining the Dispatch](#)".

Mimno, David. 'Computational Historiography: Data Mining in a Century of Classics Journals' *ACM Journal of Computing in Cultural Heritage*, to appear.

Riddell, Allen. 'Toward a Demography of Literary Forms: Building on Moretti's Graphs' *Digital Humanities* 2011.