Developing news-based Economic Policy Uncertainty index with unsupervised machine learning

Andrés Azqueta-Gavaldón

Adam Smith Business School, University of Glasgow, Room 101, 12 Southpark Terrace, Glasgow G12 8LG, United Kingdom

HIGHLIGHTS

- Employ LDA to recover topics underpinning aggregate Economic Policy Uncertainty.
- The approach economizes on costly human classification to pre-define a set of keywords.
- EPU index derived using LDA is validated with the one derived using existing methods.

ARTICLE INFO

ABSTRACT

I propose creating a news-based Economic Policy Uncertainty (EPU) index by employing an unsupervised algorithm able to deduce the subject of each article without the need for pre-labeled data.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

A novel way to compute news-based Economic Policy Uncertainty (EPU) has recently been developed by Baker et al. (2016). Their approach is based on calculating the proportion of *press articles* (*articles* from now on) describing this specific type of uncertainty over a given period of time. Nevertheless, to accurately identify those articles describing EPU, a meticulous manual process was needed. Baker et al. (2016) engaged several research assistants to manually select those articles describing EPU from a pool of 12,000 articles containing the words *economy* and *uncertainty*.[1] To be labeled as describing EPU, articles had to describe any of these eight policy categories: *fiscal* or *monetary policy*, *healthcare*, *national security*, *regulation*, *sovereign debt* & *currency crisis*, *entitlement programs* and *trade*. These articles were then used to identify and validate the combination of terms (keywords) that resulted in

the lowest gross error rate (sum of false positive and false negative selection errors).

In this note, I show how the EPU index can be built using a less costly and more flexible approach. To do so, I use an unsupervised algorithm that automatically recovers the themes of each article. The EPU index can easily be constructed by selecting only those articles whose theme matches any of the eight categories. Unlike Baker's approach, this method does not require human-classified articles in order to identify the most relevant keyword. Moreover, in addition to producing an aggregate EPU index, this method easily produces separate indices for each of the topics that underpin the EPU index.

Nevertheless, this approach does not endogenously generate the number of topics implicit in a collection of articles and must be thus set exogenously. Therefore, I use a Bayesian model selection process to find the highest *marginal likelihood* of the data (articles) when different numbers of topics are selected. The resulting index produced by aggregating only those articles describing EPU closely matches the EPU index produced using the keywords approach. The computations undertaken in this alternative process take only a few hours.

---

[1] By using any form of the terms *economy* and *uncertainty*, Baker et al. (2016) enlisted the articles describing overall economic uncertainty.

EPU has attracted substantial interest. It has been used to understand a wide range of economic and financial variables: stock prices (Pastor and Veronesi, 2012; Brogaard and Detzel, 2015); risk premia (Pastor and Veronesi, 2013); economic performance (Bachmann et al., 2013; Fernandez-Villaverde et al., 2015); corporate investment (Gulen and Ion, 2016); labor market dynamics (Bakas et al., 2016); and political polarization (Azzimonti and Talbert, 2014).

2. Latent dirichlet allocation

The Latent Dirichlet Allocation LDA model developed by Blei et al. (2003), is an unsupervised machine learning algorithm that learns the underlying topics of a set of documents (individual articles in this case). It is based on a generative probabilistic approach to infer the distribution of words that defines a topic, while simultaneously annotating articles with a distribution of topics.[2] It is an unsupervised algorithm because it *learns* these two latent (unknown) distributions of the model without the need for labeled articles.

The model recovers these two distributions by obtaining the model parameters that maximize the probability of each word appearing in each article given the total number of topics K. The probability of word w_i appearing in an article is then given by:

$$P(w_i) = \sum_{j=1}^{K} P(w_i|z_i=j) P(z_i=j) \quad (1)$$

where z_i is a latent variable indicating the topic from which the *ith* word was drawn, $P(w_i|z_i=j)$ is the probability of word w_i being drawn from topic j and $P(z_i=j)$ is the probability of drawing a word from topic j in the current article. Intuitively, $P(w|z)$ indicates which words are important to a topic, whereas $P(z)$ states which of those topics are important to an article.

The goal is therefore to maximize $P(w_i|z_i=j)$ and $P(z_i=j)$ from Eq. (1). However, a direct maximization process is susceptible of finding local maxima and showing slow convergence (Griffiths and Steyvers, 2004). To deal with this issue, I use *online variational Bayes* as proposed by Hoffman et al. (2010). This method approximates the posterior distribution of $P(w_i|z_i=j)$ and $P(z_i=j)$ using an alternative and simpler distribution: $P(z|w)$, and associate parameters.[3]

3. Material and methods

The starting point of this experiment was to download all available articles containing any form of the terms *economy* and *uncertainty* from the following newspapers: *The Washington Post*, *The New York Times*, and *USA Today*. The retrieval tool used was *Nexis*, an online database with extensive media articles coverage. The total number of articles associated with any form of these two terms from January 1989 to August 2016 was 40,454. In this *corpus*, aggregation of all articles, there are over one million unique words.

Next, the data (words) were pre-processed: *stopwords* are removed (words that do not contain informative details about an article, e.g. *at* or *and*), all words have been converted to lower case, and each word has been converted to its root (e.g. *economies* and *economist = econom*).

Finally, to find the most likely value of topics K for this specific corpus, I used the *likelihood* method. This method consists of estimating empirically the likelihood of the probability of words for a different number of topics $P(w|K)$. This probability cannot
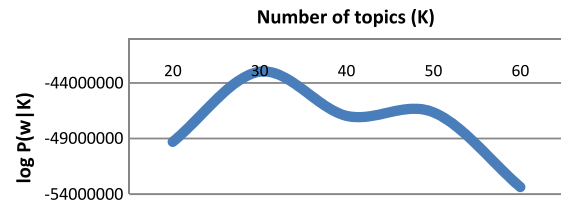

Fig. 1. Number of topics and log-likelihood scores.

Table 1
EPU categories matched by topics.

EPU subcategory	Corresponding LDA topic	Top keywords ($\lambda = 0.5$)[a]
Fiscal Policy – Taxes – Government Spending & other	Fiscal Policy	*(tax, budget, cut, bill, congress, propos, would, spend, legisl, senat, plan, fiscal)*
Monetary Policy	Monetary Policy	*(fed, economi, rate, growth, economist, inflat, econom)*
Healthcare	Healthcare	*(health, airlin, medic, patient, insur, hospit, care, doctor)*
National Security	Conflict	*(iraq, war, militari, iraqi, syria, afghanistan, attack, troop)*
	Russia	*(russia, russian, soviet, putin, ukrain, nuclear, moscow, iran)*
	Immigration	*(refuge, immigr, polici, migrant, africa, cuba, puerto, border)*
Regulation – *Financial regulation*	Law	*(court, law, legal, case, justic, rule, investig, lawyer, judg)*
	Energy	*(plant, water, energi, electr, coal, environment, farm)*
	Stock market Financial invest.	*(1,percent, 2, 3, fell, 4, rose)* *(stock, market, investor, invest, fund, yellen, wall)*
Sovereign debt & currency crisis	Financial crisis	*(bank, loan, financi, debt, credit, lender, billion, lend, default)*
	Great recession	*(bond, 2008, rate, 2012, 2013, 2011, 2014, 2016, 2009, yield)*
Entitlement Programs	Healthcare	*(health, airlin, medic, patient, insur, hospit, care, doctor)*
	Education	*(school, student, colleg, univers, educ, children)*
Trade Policy	Trade	*(china, chines, japan, india, beij, japanes, asia, taiwan, asian, currenc, trade, foreign)*

[a] The relevance of a term (w) per topic (k) is given by $(w|K) = \lambda * p(w|k) + (1 - \lambda) * p(w|k)/p(w)$, where $\lambda \in \{0, 1\}$, and $p(w)$ is the frequency of a word appearing in the corpus (see Sievert and Shirley, 2014).

be directly estimated since it requires summing over all possible assignments of words to topics, but can be approximated using the harmonic mean of a set of values of $P(w|z, K)$, when z is sampled from the posterior distribution (Griffiths and Steyvers, 2004). This method indicates that the most likely number of topics in this corpus is $K = 30$ (Fig. 1).

4. Results

4.1. Topics

Fig. 2 shows each of the 30 topics unveiled by the LDA model in this corpus. Topics are represented as circles in the two-dimensional plane whose centers are determined by computing the distance between topics (see Chuang et al., 2012).

Baker et al. (2016) identify eight categories composing EPU. These categories appear in Table 1 (column 1) together with their

---

[2] Each topic is composed of a set of most probable words while each article is composed of a set of most probable topics.

[3] For more details about the implementation see Řehůřek and Sojka (2010).