Economics Letters 158 (2017) 47–50

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

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

Article history:
Received 27 February 2017
Received in revised form 22 June 2017
Accepted 22 June 2017
Available online 27 June 2017

JEL classification:
C80
D80

Keywords:
Unsupervised machine learning
Latent Dirichlet Allocation (LDA)
Economic Policy Uncertainty (EPU)

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.

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.

E-mail address: a.azqueta-gavaldon.1@research.gla.ac.uk.

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

http://dx.doi.org/10.1016/j.econlet.2017.06.032
0165-1765/© 2017 Elsevier B.V. All rights reserved.

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:

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

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 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).

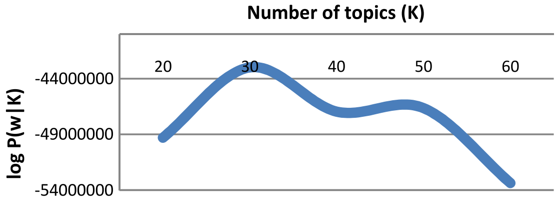

Fig. 1. Number of topics and log-likelihood scores.

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	Law	*(court, law, legal, case, justic, rule, investig, lawyer, judg)*
– Financial regulation	Energy	*(plant, water, energi, electr, coal, environment, farm)*
	Stock market	*(1,percent, 2, 3, fell, 4, rose)*
	Financial invest.	*(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).

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).

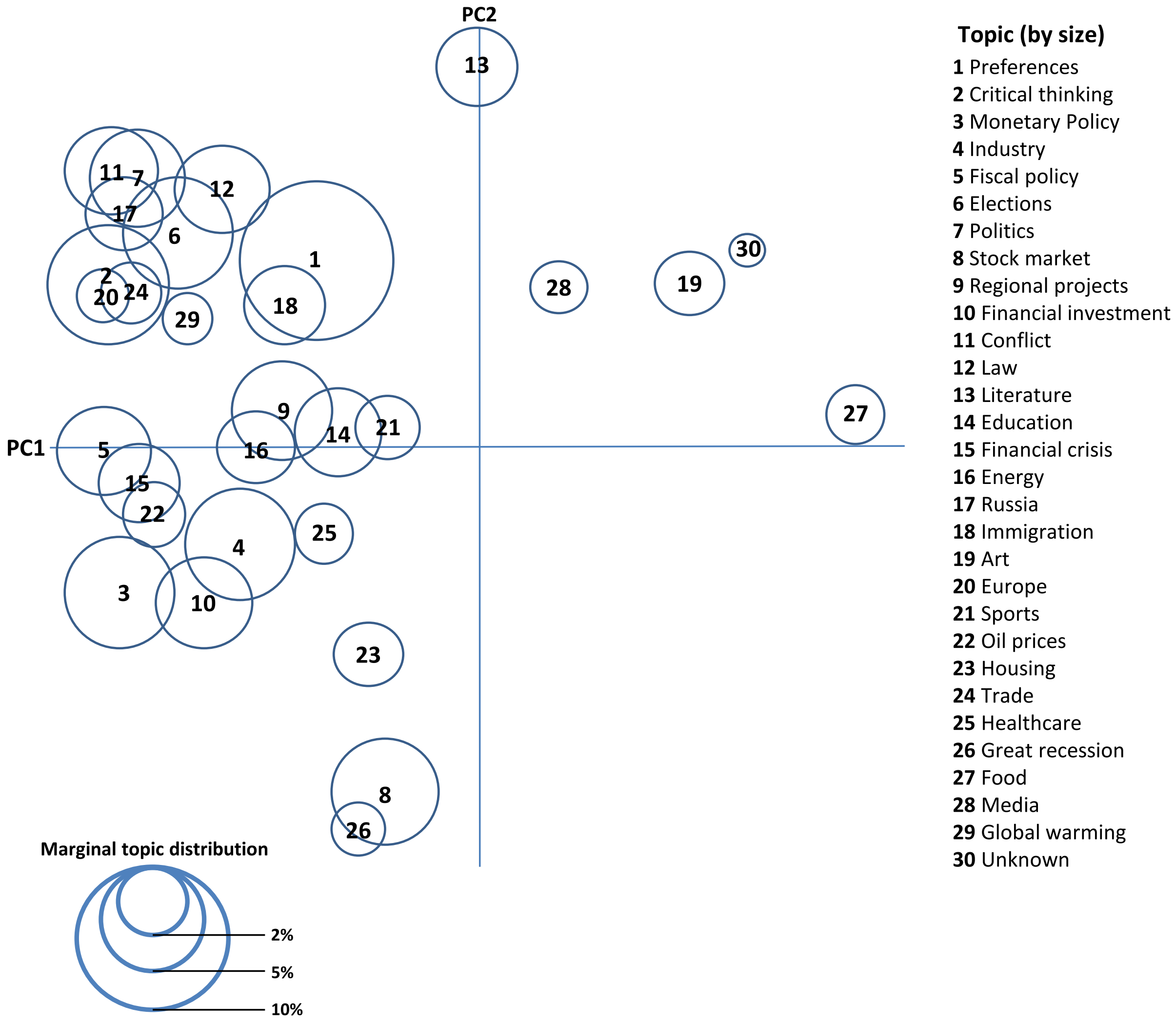

Fig. 2. Topics unveiled by the LDA. This figure is produced using the library LDAvis developed by Sievert and Shirley (2014).

equivalent topic (column 2) and the list of representative words for each topic (column 3). Although for some categories, LDA topics cannot be subdivided as suggested by those authors (e.g. *Taxes* and *Government Spending* is matched by the unique topic *Fiscal Policy*), in other cases, some topics go beyond the categories proposed by them (e.g. *National Security* can be unbundled into *Conflict*, *Russia*, and *Immigration*). Moreover, to match the category *Regulation*, I selected those topics with the highest word distribution of this term (*regulation*): *Law* and *Energy*.

Once the topics that compose EPU are found, building the EPU index required a few steps. Firstly, each article was labeled according to its most representative topic (the topic with the highest percentage in the article). Secondly, a raw count of the number of articles for every topic in each month was produced (30 *raw time-series*). Since the number of articles is not constant over time, I divided each *raw time-series* by the total number of articles containing the word *today* each month (the proxy for the total number of articles, see Azzimonti, 2015). The EPU index is then the sum of the monthly *normalized time-series* of the topics that are assigned to each EPU category. Lastly, the resulting index is standardized to mean 100 and one standard deviation. I refer to this final time-series as **EPU”**.

4.2. EPU indices

Fig. 3 shows the evolution of **EPU”** and the economic policy uncertainty index built using Baker et al.’s 2016 approach: **EPU**. This last index is produced by retrieving only those articles that satisfy the following Boolean series: {[*uncertain* OR *uncertainty*] AND [*economic* OR *economy*] AND [*regulation* OR *Federal Reserve* OR *deficit* OR *congress* OR *legislation* OR *white house*]}.[4] In order to build the final time-series, the total number of articles that contain the above set of *keywords* is divided by the total amount of articles that contain the word *today* per month, and standardize the resulting series to mean 100 and one standard deviation.

The behavior of the two time-series is extremely similar (0.94 correlation) but there are small differences regarding the intensity of some shocks. These tend to be associated with geopolitical events such as the Gulf War I, 9/11, Gulf War II and the Bush Jr. election, where the **EPU”** reacts more abruptly. These differences aside, the cyclical component between the two series is very similar (0.88 correlation), while the trend component is extremely similar (0.99 correlation).[5]

[4] Note that any form of the above list of words is retrieved. For example, legislator, legislations or legislative are forms of the word legislation.

[5] To compute the correlation between the cyclical and trend components of the two series I used the Hodrick–Prescot filter with a monthly weighted factor of 12,9600.

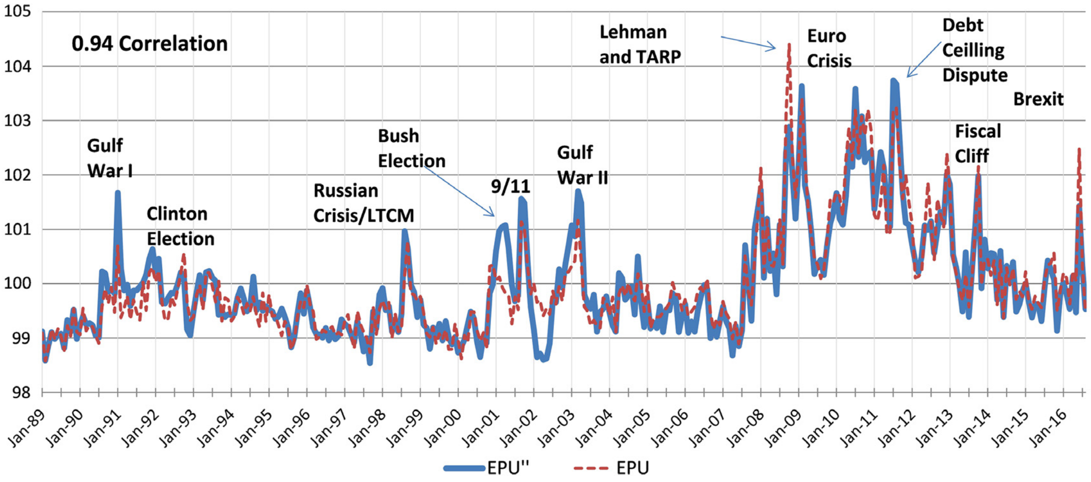

Fig. 3. Comparison between **EPU"** (solid line) and **EPU** (dashed line) indices (January 1989 – August 2016). All series are standardized to mean 100 and 1 standard deviation along the period covered.

5. Conclusion

This note shows how an EPU index may be constructed using an intuitive and quite costless approach compared to existing methods. The unsupervised nature of the model employed in this note allows classifying large textual data into topics without the need for pre-classification. The topics produced from a set of news articles describing overall economic uncertainty are easily matched with the categories that Baker et al. (2016) defined as composing EPU. The resulting index, produced within few days, greatly resembles the EPU index produced using the conventional approach which took around two years to complete.

Acknowledgments

I thank my supervisors Charles Nolan and Campbell Leith for their support and advice. Moreover, I thank the participants of the *International Conference on machine learning and computing* (ICMLC Feb 2017, Singapore) for comments and suggestions.

References

Azzimonti, M., 2015. Partisan conflict and private investment. NBER Working Paper No. 21273. http://dx.doi.org/103386/w21273.

Azzimonti, M., Talbert, M., 2014. Polarized business cycles. J. Monetary Econ. 67. http://dx.doi.org/10.1016/j.jmoneco.2014.07.001.

Bachmann, R., Elstner, S., Sims, E.R., 2013. Uncertainty and economic activity: Evidence from business survey data. Am. Econ. J.: Macroeconomics 5 (2). http://dx.doi.org/10.1257/mac.5.2.217.

Bakas, D., Panagiotidis, T., Pelloni, G., 2016. On the significance of labor reallocation for European unemployment: Evidence from a panel of 15 countries. J. Empir. Finance 39 (B). http://dx.doi.org/10.1016/j.jempfin.2016.05.003.

Baker, S., Bloom, N., David, S., 2016. Measuring economic policy uncertainty. The Q. J. Econ. 131 (4). http://dx.doi.org/10.1093/qje/qjw024.

Blei, D.M., Ng, A., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Brogaard, J., Detzel, A., 2015. The asset-pricing implications of government economic policy uncertainty. Manage. Sci. 61 (1). http://dx.doi.org/10.1287/mnsc.2014.2044.

Chuang, J., Ramage, D., Manning, C., Heer, J., 2012. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. CHI.

Fernandez-Villaverde, J., Guerron-Quintana, P., Kuester, K., Rubio-Ramirez, J., 2015. Fiscal volatility shocks and economic activity. Am. Econ. Rev. 105 (11). http://dx.doi.org/10.1257/aer.20121236.

Griffiths, T., Steyvers, M., 2004. Finding scientific topics. In: Proceedings of the National Academy of Sciences of the United States of America, PNAS, vol. 101 (Supplement 1), pp. 5228–5235. April, 2004. http://dx.doi.org/101073/pnas.0307752101.

Gulen, H., Ion, M., 2016. Policy uncertainty and corporate investment. Rev. Financ. Stud. 29 (3). http://dx.doi.org/10.2469/dig.v46.n7.12.

Hoffman, M.D., Blei, D.M., Bach, F., 2010. On-line learning for latent Dirichlet allocation. In: Neural Information Processing System.

Pastor, L., Veronesi, P., 2012. Uncertainty about government policy and stock prices. J. Finance 67 (4). http://dx.doi.org/10.1111/j.1540-6261.2012.01746.x.

Pastor, L., Veronesi, P., 2013. Political uncertainty and risk premia. J. Financ. Econ. 110 (3). http://dx.doi.org/10.3386/w17464.

Řehůřek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks. http://dx.doi.org/1013140/2123931847.

Sievert, C., Shirley, K.E., 2014. LDAvis: A method for visualizing and interpreting topics. In: ACLWorkshop onInteractiveLanguageLearning. http://dx.doi.org/1013140/2113943043.