Measure Country-Level Socio-Economic Indicators with Streaming News: An Empirical Study

Bonan Min
Raytheon BBN Technologies
bonan.min@raytheon.com

Xiaoxi Zhao
Department of Economics, Boston University
xiaoxiz@bu.edu

Abstract

Socio-economic conditions are difficult to measure. For example, the U.S. Bureau of Labor Statistics needs to conduct large-scale household surveys regularly to track the unemployment rate, an indicator widely used by economists and policy makers. We argue that events reported in streaming news can be used as "micro-sensors" for measuring socio-economic conditions. Similar to collecting surveys and then counting answers, it is possible to measure a socio-economic indicator by counting related events. In this paper, we propose Event-Centric Indicator Measure (ECIM), a novel approach to measure socio-economic indicators with events. We empirically demonstrate strong correlations between ECIM values to several representative indicators in socio-economic research.

1 Introduction

Socio-economic indicators are powerful instruments for measuring economic conditions and the sociocultural environment that people live in. They are widely used to inform policy makers, and help them to measure outcome of policy interventions. They are often difficult, if not impossible, to measure. Take unemployment rate as an example, the U.S. Bureau of Labor Statistics conducts large-scale household surveys in order to track it. More abstract indicators (e.g. economic uncertainty), which involve vague or complex social interactions, are very difficult to measure accurately.

We hypothesize that streaming news, reporting a vast amount of real-world events, can be used for measuring socio-economic indicators. We propose Event-Centric Indicator Measure (ECIM), a novel approach to measure socio-economic indicators using events extracted from streaming news. We demonstrated that ECIM is effective; ECIM values are strongly correlated with representative socio-economic indicators.

Event triggers	Location	Time	Related*
the U.S. **recession** began in December 2007	U.S. (USA)	December 2007 (2007-12)	Y
SEPT. 1, 2011 The **bankruptcies** of three American solar power companies in the last month	American (USA)	Last month (2011-08)	Y
The man accused of fatally **shooting** his estranged wife inside a New Jersey church last Tuesday	New Jersey (USA)	Last Tuesday (2009-07)	N
…	…	…	

Figure 1: The ECIM workflow. Words in bold are event triggers. Text in parentheses shows normalized locations (countries) and time. "Related" shows whether an event trigger is related to the target indicator (e.g., "unemployment rate") based on keyword matching.

We will first present an overview of the ECIM approach, and then describe how we extract events from text and aggregate them to calculate ECIMs for each socio-economic indicator. We will then present large-scale experiments to demonstrate that ECIM is effective, using several widely used socio-economic indicators.

2 Event-Centric Indicator Measure

The ECIM workflow is summarized in Figure 1. Given a large-scale streaming news collection, the system first extracts event mentions (event trigger words or phrases) along with their locations and time. For each socio-economic indicator, the system will then aggregate relevant events per each time step to produce ECIM, a time series measuring the corresponding socio-economic condition.

2.1 Extract Events from Text

We extract event frames from predicate-argument structures that are automatically generated from

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1249–1254, Hong Kong, China, November 3–7, 2019.

text. Then we extract an event trigger, a location and a time, if available, from each event frame. To attach a location or time that is further away from the trigger (or based on the publishing date), we apply a few inference rules.

Extract Events Syntactic-semantic representations such as Abstract Meaning Representation (Huang et al., 2016) and Semantic Role Labeling (SRL) (Peng et al., 2016; Surdeanu et al., 2003), have been shown to be effective for event extraction. Following (Peng et al., 2016; Surdeanu et al., 2003), our event extractor is based on tagging predicates (verbs and eventive nouns) and their locations and temporal arguments.

We first apply SRL (Punyakanok et al., 2004; He et al., 2017) over each sentence. An example SRL representation on a sentence is shown in Figure 2. From the predicate-argument structure, the system then extracts the predicate as an event trigger [1], the entity mention and time mention attached to the predicate through AM-LOC and AM-TMP as the location and time for the event respectively. The location is resolved to a country-level GPE by looking up part-of relations in GeoNames [2] (e.g., convert "Boston" into "U.S."). Our system also normalizes time mentions into Timex2 (Ferro et al., 2001) and resolves relative time (e.g., *last Tuesday*) into Timex2, based on the document publishing date. This process extracts an event mention in the form of a triple <trigger: *clashed*, location: Syria, time: 2018-11-23> from the sentence in Figure 2.

Figure 2: An SRL representation of a sentence.

Find relevant event triggers Given an indicator, we use the following approach to find a set of relevant event trigger words/phrases:

- We process a development corpus (500 documents from the English Gigaword [3]) with the above-mentioned approach to extract all event triggers, and then ask an annotator to find trigger words for the target indicator, starting with the most frequent trigger.
- We then use WordNet (Miller, 1995) synsets to automatically expand the trigger word list. We also use word embedding similarities to search for new trigger words that are most similar (e.g., using cosine similarity) to the centroid of existing triggers.

An annotator then reviews the expanded triggers and removes incorrect ones. The whole process [4] takes less than 10 minutes per indicator. Example triggers for three representative indicators are listed in Table 1.

To tag event mentions, we match the predicate to the relevant keyword list [5], constructed using the processed described above.

Infer location and time Locations are crucial for identifying whether events are related to the country of interests. Time is necessary for binning events by time steps for counting. Extracting time and locations for event triggers can be challenging if they are further apart from the trigger. Oftentimes no location or time is mentioned in the same sentence where the event is stated.

To increase the coverage of location and time for events, we apply the following inference rules:

- *Same-sentence*: if the event extractor doesn't find a location or a time, but there is one and only one location (an entity mention with type GPE or Location), and/or one and only one time mention in the same sentence, it will attach the location and/or time to the event.
- *Document metadata*: A news article often come with a publishing date and a location. When neither the extractor nor the *same-sentence* rule found a location or a time for an event, we will attach the publishing location and/or date to the event.

2.2 Measure Socio-economic Indicators

We filter events by location (target country) and then use the following steps to calculate ECIM:

- *Aggregation and counting*: We bin events by time step [7]. For each time step t and event $e \in$

[1] We expand nouns into their base noun phrases.
[2] https://www.geonames.org/
[3] https://catalog.ldc.upenn.edu/LDC2011T07
[4] We used the UI described in (Chan et al., 2019) to facilitate this process.
[5] The keyword list is available at https://github.com/BBN-E/ecim.
[6] https://en.wikipedia.org/wiki/VIX
[7] One month is used as the time step, since not all numeric indices come with higher-frequency data for validating our approach.

Indicator	Event triggers	Text
Unemployment rate	*unemploy, layoff, dismissal, lay off, economic crisis, bankrupt*	*Apple to* **lay off** *190 employees from self-driving car unit.*
CBOE Volatility Index (VIX) [6]	*sell share, liquidate, scandal, debt, loan* & (events for EPU)	*Humana executives* **sell** *nearly 74000 shares worth $22.7 million.*
Economic Policy Uncertainty (EPU)	*economic slowdown, institutional weaknesses, war, crisis, terrorist attack*	*These policies and other* **institutional weaknesses** *continue to undermine prospects for sustained economic development*

Table 1: Indicators, examples of event triggers (words or phrases) and sentences

$\mathbb{E}_i$ ($\mathbb{E}_i$ is the set of events for indicator i), we then generate count $N_{e,t}$. For each indicator i, we aggregate the counts: $\sum_{e \in \mathbb{E}_i} N_{e,t'}$.

- *Normalizing counts*: Those counts are not normalized and can be inflated due to the increasing level of media activity. To normalize, we divide them by the total number of articles published in each month M_t.

- *Smoothing*: To remove noise, we smooth the normalized counts by calculating moving averages [8] centered around each t with a window of $T = 7$ time steps [9].

In summary, ECIM for indicator i at time t is defined as

$$ECIM_{i,t} = \frac{1}{T} \sum_{t' \in [t-\frac{T}{2}, t+\frac{T}{2}]} \frac{\sum_{e \in \mathbb{E}_i} N_{e,t'}}{M_{t'}}$$

3 Experiments

We use the English Gigaword corpus [10], which consists of 5.7 million articles published from 1994 to 2010, from a wide range of sources including the New York Times, the Associated Press, Los Angeles Times, Washington Post, Agence France-Presse, Central News Agency of Taiwan, and Xinhua News Agency.

We run event extractors on this corpus to extract over 10 million event mentions that happen in the U.S. from 1994 to 2010. We then generate ECIMs for 3 representative socio-economic indicators: (1) Unemployment rate—a crucial index of economy and for policy making, (2) Chicago Board Options Exchange (CBOE) Volatility Index (VIX) —a widely-used market volatility measurement, and (3) Economic Policy Uncertainty (EPU) (Baker et al., 2016)—a policy uncertainty index. We focus on these 3 indicators because (1) they are widely used in economics and social science research, and (2) their data are publicly available [11].

Figure 3 shows the time series of ECIMs (blue solid lines) and that of the 3 corresponding indicators (red dash lines):

Unemployment rate: The downward/upward trends and the peaks of the two lines match each other quite well. This shows the ECIM is correlated with unemployment rate. There are some delay between peaks and downward trends shown in the line for unemployment rate: not surprisingly, unemployment rate reacts to events such as "economic downturn" with delay, and its recovery takes longer time than media coverage on "economic downturn" events. Among the events detected, our system found many events that can be labeled as "economic downturn" (e.g., "recession", "depression", "financial crisis") or "bankruptcy" (e.g., "bankrupt"). This matches our intuition that economic downturn and more bankruptcies are often correlated with higher unemployment rate (Reinhart and Rogoff, 2009).

VIX: Figure 3 shows that the ECIM matches very well with VIX over time. We found that high market volatility is strongly correlated with unfavorable macroeconomic events such as "economic crisis", firm-level events such as "bankruptcy", as well as its after effects such as "loan".

EPU: The ECIM strongly correlates with EPU. Similarly, economic crisis, which often led to high economic policy uncertainty, is found to be among events the most frequently detected. In addition, our system found extreme events such as "attack", "conflict", and "terrorism" which may trigger major changes of economic policy. The slight deviation in 2003-2004 is caused by low-coverage of

[8] https://en.wikipedia.org/wiki/Moving_average

[9] Other window sizes generate similar results. Due to space limitation, we only present results using 7 time steps.

[10] https://catalog.ldc.upenn.edu/LDC2011T07.

[11] We downloaded these three datasets from Federal Reserve Economic Data, CBOE, and EPU websites, respectively. For VIX which is available daily, we generated monthly averages. The cleaned version of the time-series data is available at https://github.com/BBN-E/ecim.

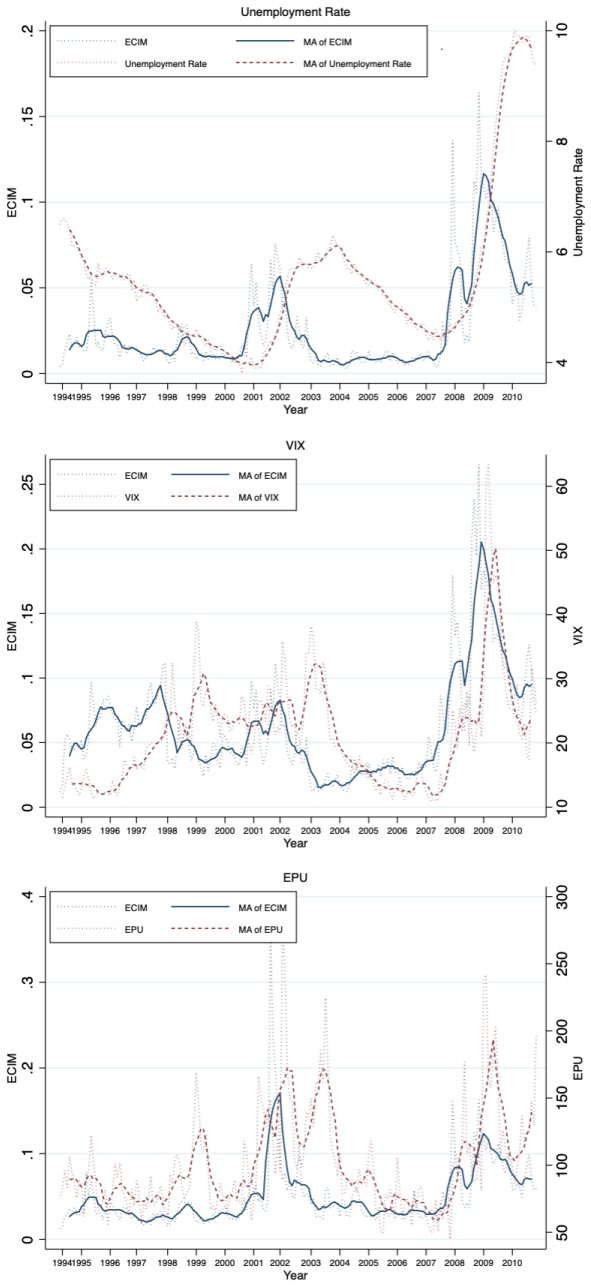

Figure 3: Values of ECIMs (blue solid lines) and the 3 corresponding socio-economic indicators (red dash lines) unemployment rate, VIX and EPU through time. Blue solid lines show ECIMs. Dotted lines show unsmoothed values. Solid and dash lines show the moving average (MA) which are smoothed values. X-axis is the year. Left Y-axis is for the ECIM value. Right Y-axis is for the indicator.

the Iraq War in Gigaword.

Table 2 shows quantitative correlation analysis between ECIMs and their corresponding indicators. Pearson correlation coefficients[12] show strong correlation between these two for each indicator. We also test the hypothesis that ECIMs and their corresponding indicators are independent (p-value is shown in the third column). For all three indicators, we are 99.99% confident to reject the hypothesis of independence.

[12] en.wikipedia.org/wiki/Pearson_correlation_coefficient

Indicator	Pearson	P-value
Unemployment rate ($1/\mathbb{M}_{comp}$)	0.4286 (0.4877)	0.0000 (0.0000)
EPU News	0.5136	0.0000
VIX	0.4115	0.0000

Table 2: Correlation coefficients between ECIMs and indicators. For unemployment rate, we also show its correlation with $1/\mathbb{M}_{comp}$ (in parentheses).

On negative correlation We further study negative correlation between some events and indicators. We hypothesize that "Competition" is positively correlated with more participants in a market, therefore a higher demand for labor which means lower unemployment rate. We construct an ECIM $\mathbb{M}_{comp}$ for "Competition" and then plot $1/\mathbb{M}_{comp}$ over time against unemployment rate. It is important to note that we use $1/\mathbb{M}_{comp}$ to flip the line for "Competition" upside down, so that the better the two lines aligns with each other, the more these two are negatively correlated.

Figure 4 shows a strikingly high negative correlation between "Competition" and unemployment rate. The correlation coefficients between unemployment rate and $1/\mathbb{M}_{comp}$ are also very high as shown in the parentheses in the first row of Table 2. This points to a promising future direction: it is possible to measure an indicator using an event that occurs more frequently when the value of the indicator is low.

4 Related Work

Most prior work on constructing or measuring indicators(Bansal et al., 2005; Jurado et al., 2015; Bachmann et al., 2013) is in social or economic science research. Recent work (Dzielinski, 2012; Alexopoulos and Cohen, 2015; Baker et al., 2016) tries to incorporate text into economic research with keyword-based approaches. Similar to ECIM, EPU (Baker et al., 2016) applies a keyword-counting approach (with keywords "economic", "policy" and "uncertainty") to measure economic policy uncertainty. In contrast, this paper uses the richer information from syntactic-semantic analysis of text. Furthermore, EPU uses

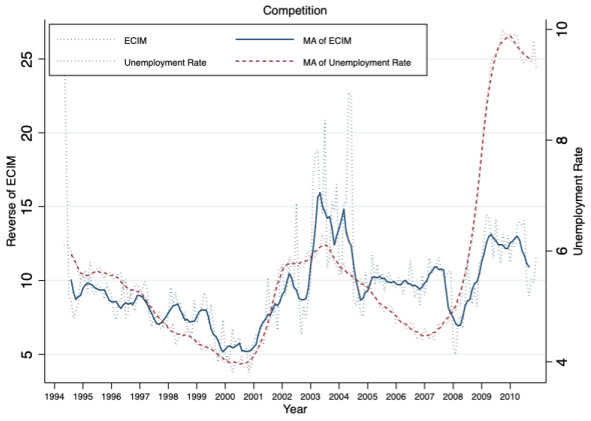

Figure 4: Values of $1/\mathbb{M}_{comp}$ (blue solid line), in which $\mathbb{M}_{comp}$ is the ECIM for "Competition", and unemployment rate (red dash line), through time.

the aggregated sentiment (by counting the number of times people expressed their views on economic policy uncertainty) as the measure, but we use a more objective approach which measures events that are correlated to the uncertainty.

(Rohlfs et al., 2016) applied supervised topic modeling to measure the effects of Federal Open Market Committee text content on the direction of short- and medium-term interest rate movements.

5 Conclusion and Future Work

This paper presents ECIM, a novel approach to measure socio-economic indicators with news events. Experiments show strong correlations between ECIM values and representative indicators in socio-economic research.

Our next steps are to further study the correlation with time lags, and to incorporate more sophisticated event extraction techniques.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments, which helped us to improve the final version of the paper.

This work was supported by DARPA/I2O and U.S. Army Research Office Contract No. W911NF-18-C-0003 under the World Modelers program. The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Department of Defense or the U.S. Government. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

References

Michelle Alexopoulos and Jon Cohen. 2015. The power of print: Uncertainty shocks, markets, and the economy. *International Review of Economics & Finance*, 40:8–28.

Rüdiger Bachmann, Steffen Elstner, and Eric R Sims. 2013. Uncertainty and economic activity: Evidence from business survey data. *American Economic Journal: Macroeconomics*, 5(2):217–49.

Scott R Baker, Nicholas Bloom, and Steven J Davis. 2016. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Ravi Bansal, Varoujan Khatchatrian, and Amir Yaron. 2005. Interpretable asset markets? *European Economic Review*, 49(3):531–560.

Yee Seng Chan, Joshua Fasching, Haoling Qiu, and Bonan Min. 2019. Rapid customization for event extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 31–36, Florence, Italy. Association for Computational Linguistics.

Michal Dzielinski. 2012. Measuring economic uncertainty and its impact on the stock market. *Finance Research Letters*, 9(3):167–175.

Lisa Ferro, Inderjeet Mani, Beth Sundheim, and George Wilson. 2001. Tides temporal annotation guidelines version 1.0. 2. *The MITRE Corporation, McLean-VG-USA*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 258–268.

Kyle Jurado, Sydney C. Ludvigson, and Serena Ng. 2015. Measuring Uncertainty. *American Economic Review*, 105(3):1177–1216.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Haoruo Peng, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 392–402.

Vasin Punyakanok, Dan Roth, Wen-tau Yih, and Dav Zimak. 2004. Semantic role labeling via integer linear programming inference. Technical report, ILLINOIS UNIV AT URBANA-CHAMPAIGN DEPT OF COMPUTER SCIENCE.

Carmen M Reinhart and Kenneth S Rogoff. 2009. The aftermath of financial crises. *American Economic Review*, 99(2):466–72.

Christopher Rohlfs, Sunandan Chakraborty, and Lakshminarayanan Subramanian. 2016. The effects of the content of fomc communications on us treasury rates. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2096–2102.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.