CAFRAL
CENTRE FOR ADVANCED FINANCIAL RESEARCH AND LEARNING
Promoted by Reserve Bank of India

Economic Policy Uncertainty Index : Augmentation and Segregation

Dr N R Prabhala
Raghav Somani, Sumit Karvade & Vinod Dharmarajan

Contents

1 Abstract

Uncertainty regarding economic policy has become an important factor in economic growth, crisis and recovery. It is now possible to quantify the same by combining various methods like keyword search in news articles, the dispersion in forecasts etc. We extend the Baker, Bloom and Davis (2013) methodology using Support Vector Machines to classify articles and improve precision. We also identify whether it is political or economic factors that cause policy uncertainty. We find that our uncertainty index is able to improve upon the index constructed by Baker, Bloom and Davis (2013).

2 Introduction

Baker, Bloom and Davis (2013) can be considered to be the seminal work to quantify uncertainty. Alesina et al. (1993) found that socio-political uncertainty (caused by income inequality) decreases investments. Johannsen (2014) found that fiscal policy uncertainty (in government spending and tax rates) can have large negative impact on consumption, investment and output when the zero lower bound holds, but the effects are moderate when the monetary authority is not constrained by the zero lower bound.
However these studies didn't quantify uncertainty and hence were unable to separate out the effects of policy uncertainty from other factors. Policy uncertainty has come into the limelight as it plays an important role in determining the effectiveness of various fiscal and monetary tools. Aastveit, Gisle and Sola (2013) showed that the effect of monetary policy is considerably weaker when uncertainty is high. Various studies have found economic policy uncertainty to be negatively correlated to investment and employment in the economy. Baker, Bloom and Davis (2013) find that firms with greater exposure to government spending have statistically significant negative relationship between investment and employment to the levels of policy uncertainty. Bhagat, Ghosh and Rangan (2013) find that if the policy uncertainty index was at the same level as of 2006, then Indian GDP would have grown by 0.56% and fixed investment would have increased by 1.36%. Durnev (2010) found that there is a significant decrease in investment-to-price sensitivity in an election year compared to non-election years from a cross section of 79 countries for the period 1980-2006.

Given the important impact of uncertainty on economic outcomes, we set out to examine the existing methodology in quantifying the same, improve upon it and apply it in the context of the Indian economy.

Economic Policy Uncertainty Index - Baker et al. (2013)
The index consists of three components -

1. **News coverage** - Coverage of policy related uncertainty newspapers.
   For constructing the US EPU, they picked ten largest newspaper for the period 1985-2009 and searched each article for the following set of terms - {economic or economy} & {uncertain or uncertainty} & {congress or deficit or Federal Reserve or legislation or regulation or regulatory or Fed or White House} Hence each article representing economic policy uncertainty necessarily should have the words economy/economic and uncertain/uncertainty and any one of the words from the third component. After identifying the articles and counting their number for each newspaper in a given month, they are normalized by dividing them with the total number of articles published in the same newspaper in the given month. The resulting monthly series is normalized to unit standard deviation (for the entire sample period) and then summed across the 10 newspapers.

2. **Scheduled Tax Expirations** - The second component of the EPU index is the absolute dollar value of expiring tax provisions in each year over the next 10 years. The absolute

value is discounted at the rate of 50% and then summed to obtain the index value for each January. This value is held constant for the calendar year.
The above component is not part of the EPU constructed for economies other than that of US.

3. **Forecasters disagreement** - Survey of Professional Forecasters, conducted by the Federal Reserve Bank of Philadelphia is used to measure dispersion of expectations on inflation and government spending (Federal, State and Local). The survey is conducted on a quarterly basis, initiated at the end of the first month in a quarter. The forecast dispersion thus derived is assigned to the second and third month of the given quarter and the first month of the next quarter. Forecast data on inflation is annualized Quarter on Quarter CPI rate over the next 4 quarters. The interquartile range of forecasts is used to measure the dispersion. In case of Federal purchases, the interquartile range of forecasts is first divided by its median and then multiplied by 5 year moving average of the ratio of nominal Federal purchases to nominal GDP. The same is done for State and Local government purchases. The two measures - Federal and State and Local government purchases - are then summed to measure forecaster disagreement about future government purchases expressed as a percentage of GDP.

Note that (1) has been standardized. (2) and (3) are also standardized with respect to their own standard deviations, following which all three components are added for each month with the following weights - $\frac{1}{2}$ for (1), $\frac{1}{6}$ for (2), $\frac{1}{6}$ for Inflation forecast dispersion measure and $\frac{1}{6}$ for government purchases (Federal, State and Local) forecast dispersion measure. The overall index is then rescaled such that the average value of the index is 100 over the sample period 1985-2000.

EPU for India, Bhagat et al. (2013)
They use two components -

1. **Newspaper coverage of policy related economic uncertainty** - To create a measure of the first component the following newspapers were used - The Economic Times, The Times of India, The Hindustan Times, The Hindu, The Statesman, The Indian Express and The Financial Express. Articles are first searched for terms - {uncertain or uncertainty or worry or fear} & {economic or economy} & {regulation or central bank or monetary policy or policy makers or deficit or legislation or fiscal policy}.
The number of articles containing the above search terms in each newspaper each month are then scaled by total number of articles published by that newspaper in the respective month. The series for each newspaper is standardized and then summed across. The sum is then rescaled such that the average of the series is 100.

2. **Disagreement among forecasters** - To determine dispersion over forecasts from Consensus Economics which provides forecasts on various economic variables by professional forecasters on a monthly basis. Forecasts on inflation and federal government budget balance is used to measure disagreement among forecasters. Each forecast pertains to the corresponding month in the following year. In case of inflation, the forecasts are for the consumer price index and dispersion is measured by the inter quartile range of forecasts each month. Similarly, the interquartile range of forecasts for budget balance divided by contemporaneous annual GDP is used to measure dispersion in the budget balance component. Both these measures are corrected for monthly fixed effects. In cases where forecasters report only once a quarter, the data from most recent forecast has been used for upto 3 months.

The two components - Newspaper coverage and forecasters disagreement - is first normalized by their respective standard deviation and then combined with $\frac{2}{3}^{rd}$ weight attached to the former

and $\frac{1}{6}^{th}$ each to the two subparts of the latter component.
Baker et al. (2013) however does not find increases in their EPU index around political events and assume that political uncertainty doesn't raise economic policy uncertainty.

EPU for Belgium, Tobback, Daelemans, Fortuny, Naudts and Martens (2014)
They constructed an EPU index for Belgium with the news coverage and forecasters disagreement components. However to construct the news coverage index, they used two other algorithms - Modality annotation and a Support Vector Machines (SVM) classification model.

1. **Modality annotation** - Modality annotation counts the use of words expressing uncertainty. First they constructed a list of modal items expressing uncertainty in Dutch. Then for each word in the article they computed the modality score, i.e. the relative frequency with which words in the uncertainty word list occurred. The modality scores are used to classify the articles whether they address economic policy uncertainty or not. They set the classification threshold at 15% and the relevant number of articles in a month was divided over by the number of news sources in that month.

2. **SVM classification** - Support Vector Machines on the other hand is a trained classifier, used to predict whether an article represents uncertainty or not. The algorithm is designed to look for the most discriminative words within articles to classify them. It looks for patterns in the text and selects words with largest discriminative power. They first identified whether an article represents uncertainty or not with for the articles in their training set based on which the algorithm labels news articles in the data set. It is possible that there will be misclassified articles. Such articles were then relabeled manually and then fed back as input. Thus the active learning process helps improve the definition of decision boundary. SVM algorithm attempts to define the decision boundary which will maximize the margin between the two classes.

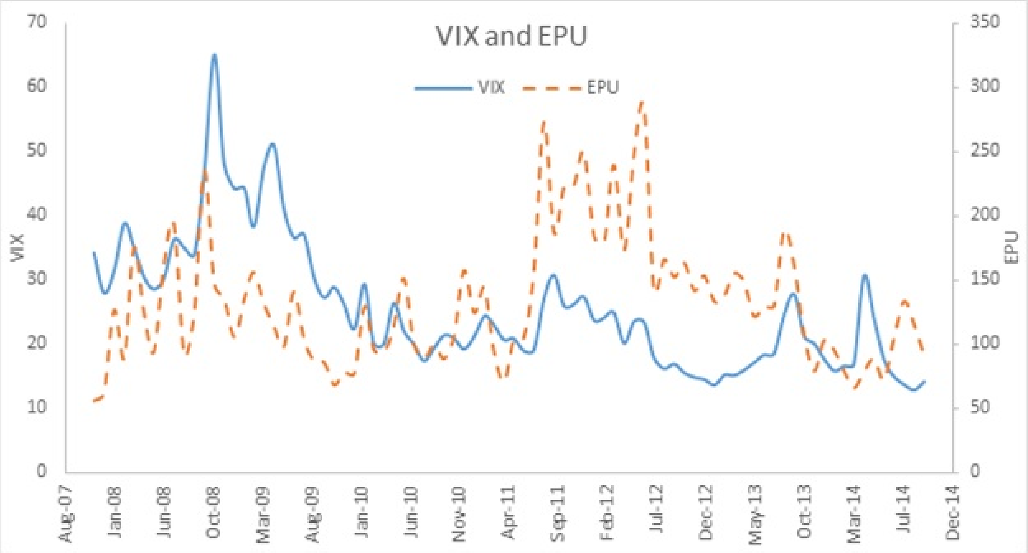
They found that false negative rates, i.e. the probability that an articles is incorrectly classified as EPU = 0 to be highest for the method used by Baker et al. (2013), the rate being 0.94 while the rate of false negatives using Modality annotation and SVM classification were 0.76 and 0.24 respectively. The stark difference in the rates could be due to the self-selection of the list of words used for classification in the method used by Baker et al. (2013) and Modality annotation. It is possible that the words "uncertainty" and "uncertain" fails to accurately capture policy uncertainty in all cases.

In case of the EPU index created for India, the only one of the seven newspapers used are among the top 10 newspapers in terms of readership. The ten biggest newspapers in terms of daily circulation[1] are:

1. The Times of India (English)
2. Dainik Bhaskar (Hindi)
3. Dainik Jagran (Hindi)
4. Hindustan (Hindi)
5. Malayala Manorama (Malayalam)
6. Amar Ujala (Hindi)
7. Eenadu (Telugu)
8. Rajasthan Patrika (Hindi)
9. Daily Thanthi (Tamil)
10. Lokmat (Marathi)

A similar story unfolds while looking at state-wise readership numbers with regional newspapers dominating the circulation charts. Moreover, it is the case that the regional newspapers have their own news bureau and reporters. Thus a strong case can be made for searching examining news coverage in the regional languages rather than focusing on English newspapers only. Examining language based newspapers provides the added advantage of deciphering regional political events and uncertainty.

Baker, Bloom and Davis (2013) also found that their EPU index is more strongly correlated with implied volatility index as the time horizon of the latter increases. For instance, the correlation between EPU and 1-month implied volatility index from VIX index is 0.578 as against 0.855 with the 10 year implied volatility index.



The EPU index for India seems to match with the NIFTY VIX for the period November, 2007 to October, 2014. However EPU remains elevated even when VIX has fallen in the period June, 2011 to May, 2013. It could imply that while overall economic uncertainty captured by VIX had decreased, policy uncertainty remained elevated.

Baker, Bloom and Davis (2013) in their study of the EPU find that firms with higher levels of exposure to government contracting and spending have significantly higher negative effects due to increase in EPU compared to firms with lower exposure.

There are various empirical studies that negative impact of political uncertainty on economic outcomes. Aisen and Veiga (2011) in their study of 169 countries using a system GMM for the period of 1960-2004 find that high political instability is associated with a lower per capita GDP growth rate. They define political instability with the variable cabinet change, which comprises of the number of times in a year a new premier is named and/or 50% or more of the cabinet positions are taken by new people.

Zarnowitz (1992) used data from surveys conducted since 1968 by American Statistical Association (ASA) and the National Bureau of Economic research (NBER). These surveys consist of numerical responses by 80 regular respondents (professional forecasters) on many variables including GNP, RGNP (in constant dollar) and IPD (GNP implicit price deflater).

Consensus is defined as the degree of agreement on point forecasts and uncertainty is defined as the degree of diffuseness of the probability distributions given by the same forecasters to their predictions. Based on analysis from the data, mean point forecasts and mean probability forecasts for the same individuals agree closely. However probability distributions show more uncertainty compared to the point estimates, especially for the short horizons (shorter term predictions). In the longer term, standard deviation for point forecasts shows more volatility than standard deviation for mean of probability distributions.Increase in inflation forecasts increases the uncertainty (mean of probability distributions). Also, increase in uncertainty is shown to have negative effects on Real GNP growth.

This provides clear evidence for the use of forecasters dispersion variables as a proxy for uncertainty in the economy as well as the negative effects of the rise of uncertainty.

Our index aims to separate out the political uncertainty, uncertainty in general, Economic uncertainty and attempts to analyze individual impacts of these factors. For example, direct fiscal policy uncertainty due to a change in the government objectives or new schemes is different from indirect fiscal policy uncertainty due to changes in the political environment, coalition government problems or death of prominent leaders in the fact that the former can be controlled but the latter cannot (in a democracy).

3 Data

For this paper, we took the following number of articles from the business and national section of 'The Hindu' newspaper

1. 6,189 articles from year 2001,
2. 13,102 articles from year 2008,
3. 13,255 articles from year 2009 and
4. 16,429 articles from year 2010.

Only two of the sections of the newspaper were selected as they were the most relevant sections for the study of political and economic factors, along with uncertainty in general. The 'Front page' section was not chosen and instead the National section was taken to avoid the inclusion of international events which do not have a significant enough impact on India that they be included in the National section. The year 2001 was chosen as it was not a good year in recent history of India, a witness to Gujarat earthquakes, WTO attack fear and subsequent financial crisis, followed by an attack on India's Parliament. The years 2008-09 were then chosen to analyze the effects of the global financial crisis in an event by event study. Further studies could as well include the rest of the years from 2001 to 2015 and make a continuous index based on the same methodology.

4 Methodology

The idea used to construct the uncertainty indices is to classify newspaper articles based on a subjective perception of uncertainty. The news articles once classified can be used to get estimates of the levels of uncertainty over a range of time.

4.1 Text Classification

As our data is large, text classification is important to organize and analyze textual data. Human categorization is very time-consuming and costly, thus limiting its applicability especially for large or rapidly changing collections. Here is where *'Machine learning'* comes into the picture.
In particular, "Machine learning is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty."[1]
Text classification can be treated as a supervised machine learning method of inferring the classification function from labeled training data.
A growing number of statistical classification and machine learning techniques have been applied to text classification, including multivariate regression, nearest neighbor classifiers, probabilistic Bayesian models, boosted decision trees, neural networks, symbolic rule learning, multiplicative update algorithms, support vector machines, regularized logistic regression and random forests.[3]

4.2 Why Support Vector Machines?

Support vector machines are based on the *Structural Risk Minimization* principle from computational learning theory [Vapnik 1995]. The idea of structural risk minimization is to find a hypothesis h for which we can guarantee the lowest true error. The true error of h is the probability that h will make an error on an unseen and randomly selected test example.
To find out a promising classifier we first need to know more about the properties of text.

- **High Dimension input space** - Our inputs are text documents and so we have a large number of features (words) to deal with. As SVM uses over-fitting protection which do not depend much on the number of features, they have a good potential to handle a large feature space. SVMs measure the complexity of hypothesis based on the margin with which they separate the data and not the number of features. This means that we can generalize even in the presence of many features, if our data is separable with a wide margin using functions from the hypothesis set.

- **Non-linearly separable data** - It is very likely to have a non-linearly separable data as the training data might be labeled using human subjective thinking and perception. SVMs can transform the dataset into a non-linear feature space (of a different dimension may be) without even actually visiting that space and gain computational advantage of it which is one of the most exciting properties of SVMs.

Support Vector Machines (SVMs) have been proven as one of the most powerful learning algorithms for text categorization.[3]

4.3 Representing Text Documents

Documents, which typically are strings of characters, have to be transformed into a representation suitable for the learning algorithm and the classification task. The representation of data has a strong impact on the generalization accuracy of a learning system. In order to represent data to have minimum out-of-sample misclassification we do the following -

1. All the non-alphabetic characters are removed.

[1] Machine Learning - A Probabilistic Perspective, Kevin P. Murphy
[3] Using SVMs for Text categorization, Susan Dumais, Microsoft Research
[3] Thorsten Joachims: Text categorization with support vector machines: learning with many relevant features, 1998

2. To get rid of different grammatical forms of a word we lemmatize words on the basis of adjective, noun, adverb and verb.
   For example - A sentence like "cats running ran cactus cactuses cacti community communities" will be lemmatized to "cat run run cactus cactus cactus community community".

3. The words which appear just once in the article set are removed to have a better generalization and thus eliminate noisy dimensions.

4. Based on this basic representation it is known that scaling the dimensions of the feature vector with their *Inverse Document Frequency IDF*(w_i) [Salton and Buckley, 1988] leads to an improved performance. IDF(w_i) can be calculated from the document frequency DF(w_i), which is the number of documents the word w_i occurs in.

   $$IDF(w_i) = \log \frac{n}{DF(w_i)} \tag{1}$$

   Here, n is the total number of training documents. Intuitively, the inverse document frequency of a word is low if it occurs in many documents and is high if the word occurs in a few.

5. An optimal number of features are selected from the training set based on absolute linear correlation with the output. To select a subset of f features, the f words with the highest absolute linear correlation with labeled output are chosen. All other words are ignored.

6. Each document feature vector $\vec{d_i}$ is normalized to unit length in l_2 space.

4.4 Support Vector Machines

SVMs are one of the most popular techniques in supervised machine learning. Linear SVM models find the best hyperplane that can separate the points of different classes with the maximum margin. As our input feature space might not have linearly separable points, we use a regularized soft margin technique to obtain the hyper-surface allowing some points to cross the margin with a limited constrain. To make use of the most interesting feature of SVM we use the Radial Basis function kernel to obtain the best soft margin hyper-surface as a decision boundary for classification.

Being a supervised learning problem, we need to train the machine using a uniform sample of training data set already labeled.

Let the training data set be

$\{ \ (\vec{x_1}, y_1), \ (\vec{x_2}, y_2) \ ..., (\vec{x_N}, y_N) \ \}; \ \vec{x_i} \ \epsilon \ \mathbb{R}^f, \ y_i \ \epsilon \ \{-1, 1\}$

where $\vec{x_i}$ represents the input feature vector of the i^{th} data point and y_i is its labeled class.

We define a hyperplane by $\quad \{ \ x : f(\vec{x}) \ = \ \vec{x}.\vec{w} + w_0 \ = \ 0 \ \}$.

We normalize $\vec{w}$ as $|\vec{w}.\vec{x}| = 1$ and pull out w_0, the constant term from $\vec{w}$ to get $|\vec{w}.\vec{x} + w_0| = 1$

The classification is induced by $\quad G(\vec{x}) = sgn(f(\vec{x}))$.

We need to know the coefficients of the plane to maximize the margin between points if the two classes are linearly separable. The margin between $f(\vec{x}) = 0$ and $|f(\vec{x})| = 1$ is $M = \frac{1}{||\vec{w}||}$.

The optimization problem is to maximize M or

$$Minimize \quad \frac{1}{2}||\vec{w}||^2 \tag{2}$$

subject to constrains

$$y_i(\vec{w}.\vec{x_i} + w_0) \geq 1 \quad \forall \ i = 1, 2, ..., N \tag{3}$$

As classes might overlap, the dataset might not be linearly separable. One way is to allow some points to cross the margin. We define slack variables $\vec{\xi} = (\xi_1, \xi_2, ..., \xi_N)$ and modify the constrains to

$$y_i(\vec{w}.\vec{x_i} + w_0) \geq 1 - \xi i \quad \forall\ i = 1, 2, ..., N \tag{4}$$

allowing the i^{th} point to violate ξ_i fraction of margin M. Therefore $\xi_i \geq 0\ \forall\ i = 1, 2, .., N$. Also we do not allow the total violation to cross a barrier C i.e.,

$$\sum_{i=1}^{N} \xi_i \leq C \tag{5}$$

The optimizing function changes to

$$Minimize_{\vec{w},b,\vec{\xi}}\ \frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{N}\xi_i \tag{6}$$

Subject to

$$y_i(\vec{w}.\vec{x_i} + w_0) \geq 1 - \xi i, \quad i = 1, 2, ..., N \tag{7}$$

and

$$\xi_i \geq 0, \quad i = 1, 2, ..., N \tag{8}$$

If the penalty C is large, most of the ξ_i tend to 0.
The Lagrange primal function of the optimization problem is

$$L_P = \frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(\vec{w}.\vec{x_i} + w_0) - (1 - \xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i \tag{9}$$

which we minimize w.r.t. $\vec{w}$, w_0 and ξ_i. By setting the respective derivatives to 0 and substituting the results back to Equation (9) we obtain the Lagrange (Wolfe) dual objective function

$$L_D = \sum_{i=0}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \vec{x_i}.\vec{x_j} \tag{10}$$

which gives a lower bound on the objective function for any feasible point. We maximize L_D in addition to the KKT constrains using standard techniques. Together the equations uniquely characterize the solution to primal and dual problem and we see that the solution of $\vec{w}$ has a form of

$$\vec{w} = \sum_{\hat{\alpha_i}>0}\hat{\alpha_i}y_i.\vec{x_i} \tag{11}$$

with non zero coefficients $\hat{\alpha_i}$ only for those observations i for which the constraints are exactly met. These observations are called *Support Vectors* since $\vec{w}$ is represented in terms of them alone. Among them some will lie on the edge of the margin and can be used to solve for w_0. Given the solution the decision function can be written as

$$\hat{G}(\vec{x}) = sign[\hat{f}(\vec{x})] \tag{12}$$

The SVM model described so far finds the linear boundaries in the input feature space. Generally linear boundaries in an enlarged space achieve better training-class separation, and translate to non-linear boundaries in the original space. Support Vector Machines can enlarge the space to get a very large or even infinite dimensional space. This is because the solution to the plane

uses only the inner product of two vectors in the space. If we can formulate the inner product of the transformed space $K(\vec{x_i}, \vec{x_j})$ and plug it in the solution we save a lot of computation cost by not transforming the whole dataset to the transformed space and performing the Linear SVM model in that space.
Here we come across kernels which are inner product functions in a transformed space.
The solution hence can be written as

$$\hat{f}(\vec{x}) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(\vec{x}, \vec{x_i}) + w_0 \tag{13}$$

Three popular choices of Kernels in SVM literature are

1. d^{th} Degree polynomial $= (1 + \vec{x_i}.\vec{x_j})^d$
2. Radial Basis $= e^{-\gamma||\vec{x_i} - \vec{x_j}||^2}$
3. Neural Network $= \tanh(\kappa_1 \vec{x_i}.\vec{x_j} + \kappa_2)$

4.5 Training the SVM

SVM being a supervised learning model requires a training set. We manually labeled some articles to get a generalized SVM model to predict the remaining leftover articles. We chose the training set in the below mentioned proportions.

Table 1: Training set

Year	Number of articles	Number of Training articles
2001	6,189	500
2008	13,102	300
2009	13,255	300

The indices for 2001 was constructed using the training data of 2001 only.
The indices for 2008 and 2009 was constructed using the training data of 2001, 2008 an 2009.
The indeces for 2010 was constructed using the training data of 2001, 2008 and 2009 only.

For training the articles at first the Uncertainty (U) is labeled. If an article is labelled as U=1 we further label it according to Economic Uncertainty (EU) and Political Uncertainty (PU). If the Uncertainty (U) is labelled as 0, we do not attempt to label any of them further. If an article is labeled as EU = 1, we label the Economic Policy Uncertainty (EPU). If EU=0, we do not attempt to label EPU.
Refer to the 'Coding guide' document for a description of the coding scheme.

4.6 SVM model selection

For the classification algorithm to work at its best, we select the best parameters to prepare the model from the labeled training set. The SVM model chosen uses the soft-margin method on Radial-Basis Kernel to gain the best generalization advantage. The parameters affecting the soft margin is the cost C (Equation 5) and the parameter affecting the Kernel is γ, the term in the exponent of the radial basis function. Also we select the terms on the basis of their absolute liner correlation with the predicted output.
For selecting the number of features (words) using linear correlation method we perform repeated 5-fold cross-validation to get a near estimate of the mean out-of-sample misclassification error over 30 permutations of trainset and testset amongst the labeled articles. We use similar

cross-validation method to find out the best model by iterating over different values of C and γ.

Parameter and feature selection was done on the basis of the obtained graphs as given in the Appendix.

Table 2: Model selection 2001

Index	Number of features	γ	C	Misclassification
U	1500	0.0006	1000	0.159
EU	2000	0.0008	600	0.12
PU	1500	0.0011	4000	0.116
EPU	1500	0.0014	3000	0.116

Table 3: Model selection 2001, 2008 and 2009

Index	Number of features	γ	C	Misclassification
U	3000	0.0008	500	0.162
EU	2500	0.0012	300	0.089
PU	2500	0.0008	600	0.1297
EPU	1500	0.0006	1000	0.1242

4.7 Constructing the Index

Using the models selected we obtain the fraction of article predicted as U=1 for constructing the Uncertainty Index (U) for a month. To obtain the conditional Economic Uncertainty Index (EU) and Political Uncertainty Index we test on only those articles which were predicted as U=1 and calculate the mean only on those articles. To obtain the conditional EPU Index we predict only the articles which were predicted as EU=1 and calculate the mean only on those articles.

The time series of the fractions obtained is normalized to mean 100 and unit standard deviation to obtain the corresponding indices.

5 Results

The normalized Indexes are plotted along the time axis.
From the below plots for 2001 we find that the Index peaks at events like Union Budget, state elections, Agra summit and 9/11 attacks.

Figure 1: 2001 U Index

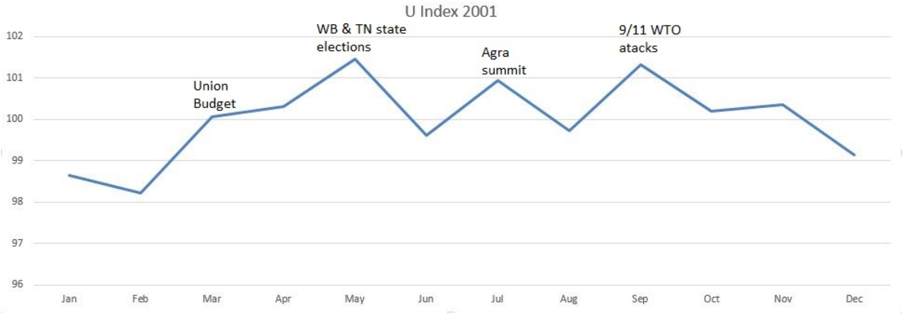


Figure 2: 2001 EU Index

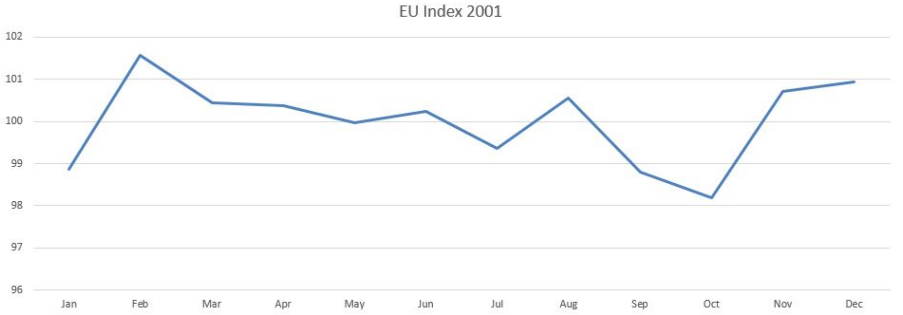


Figure 3: 2001 PU Index

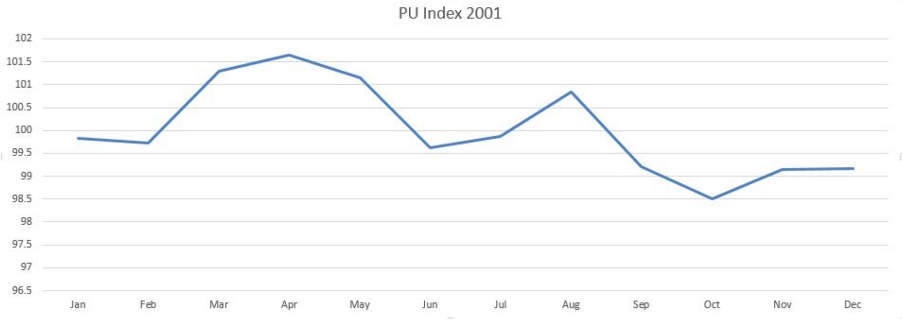
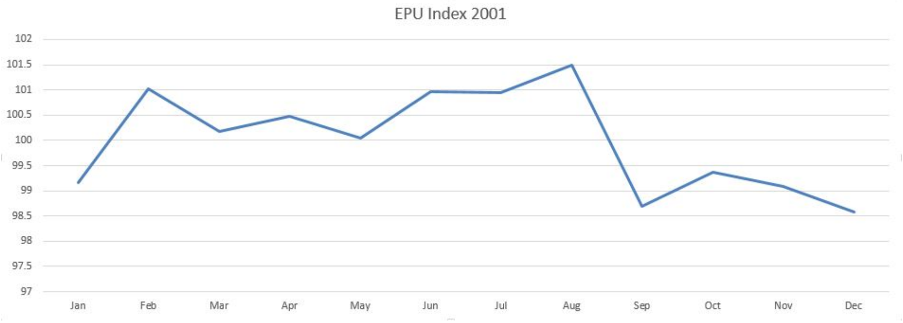


Figure 4: 2001 EPU Index



From the below plots for 2008 and 2009 we find that the index peaks at events like Global financial crisis, no confidence nuclear deal with US, Lehman bankruptcy, 2009 General Elections.

Figure 5: 2008-2009 U Index

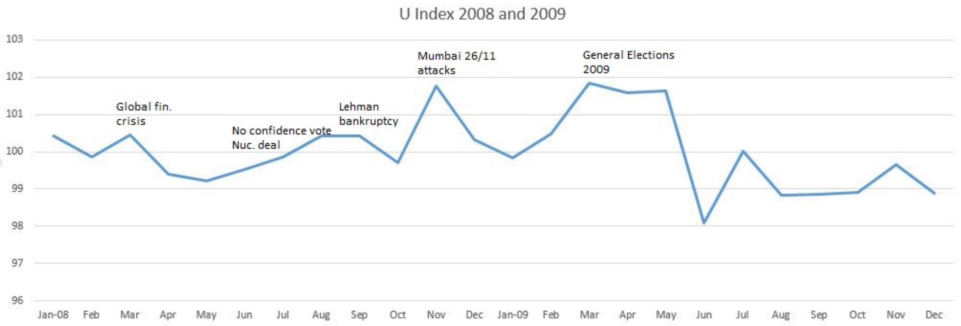


Figure 6: 2008-2009 EU Index

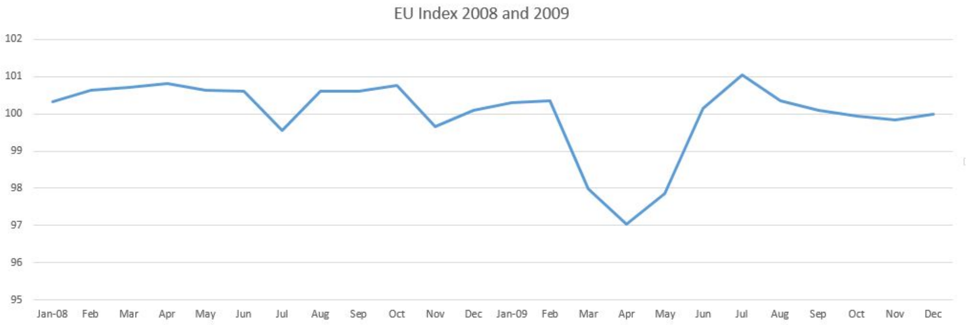


Figure 7: 2008-2009 PU Index

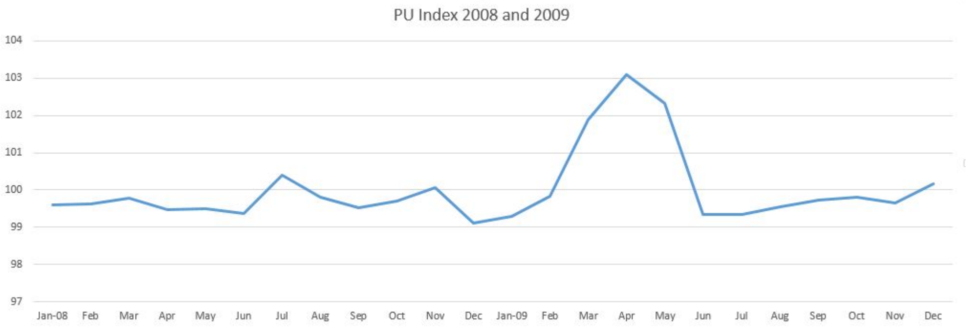
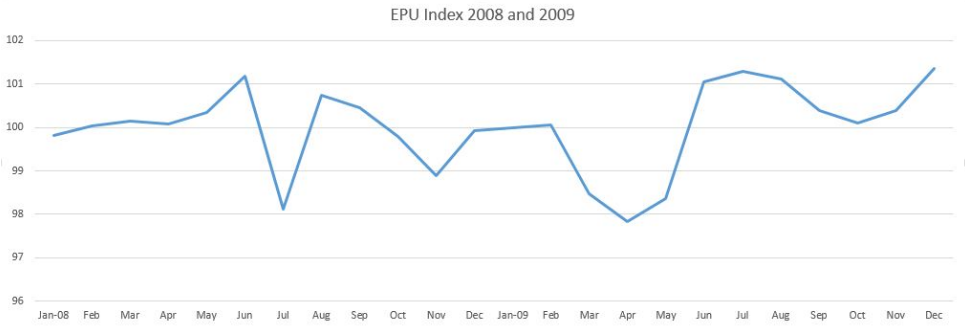


Figure 8: 2008-2009 EPU Index



From the below plots for we find that the index peaks at Union budget, Commonwealth game scam, RBI interest rate hikes.

Figure 9: 2010 U Index

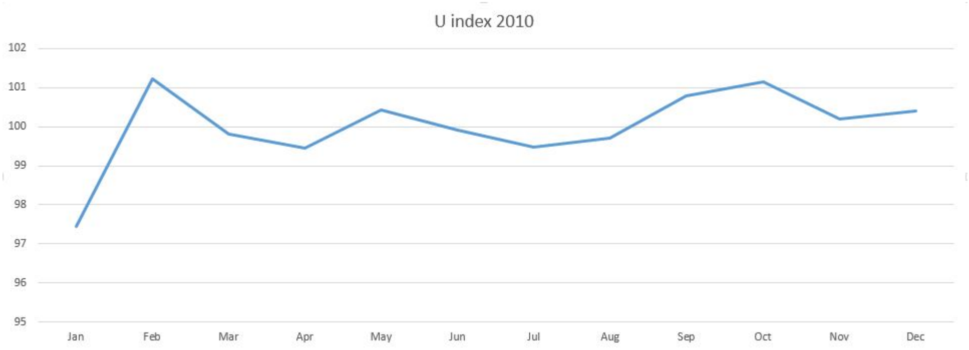


Figure 10: 2010 EU Index

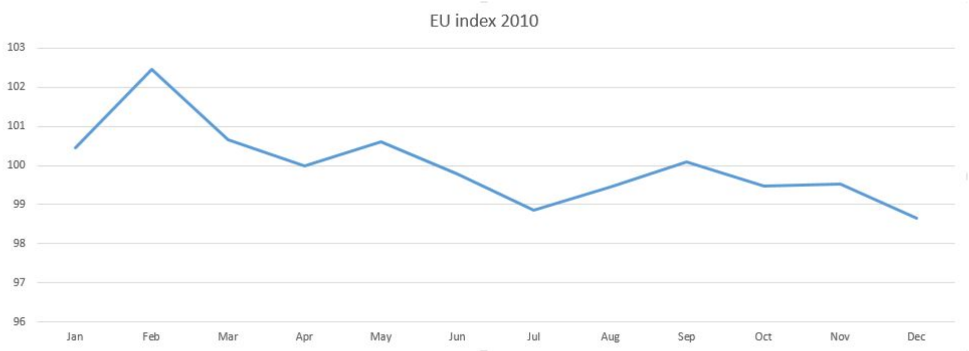


Figure 11: 2010 PU Index

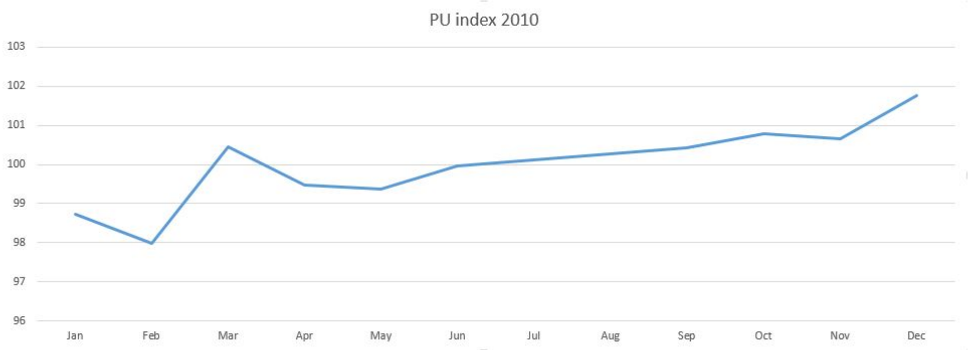
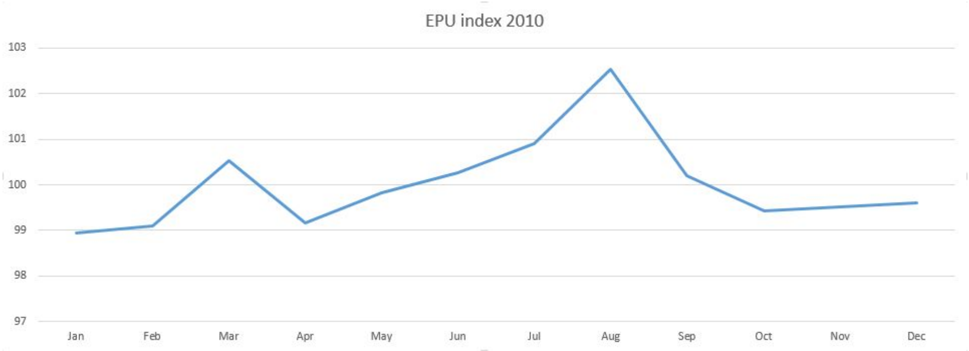
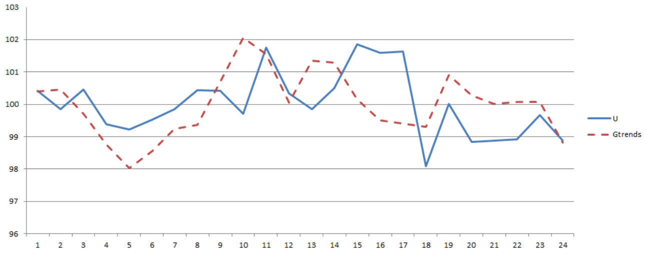


Figure 12: 2010 EPU Index

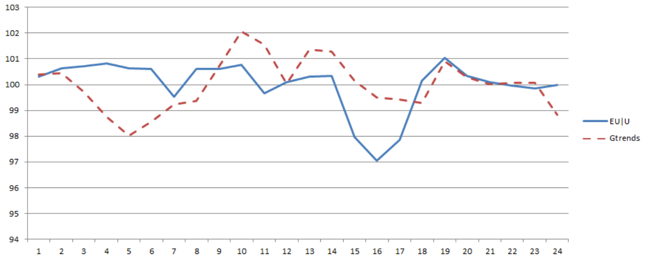


6 Validation

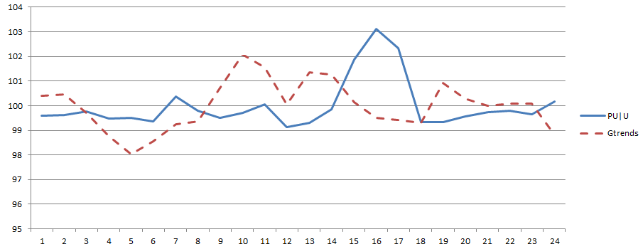
For the validation of our index, we use Google trend data for words like ‘economic’, ‘deficit’, ‘parliament’, ‘legislation’ and ‘RBI’, Indian VIX and prices of 10 year government bonds for the years 2008 and 2009.



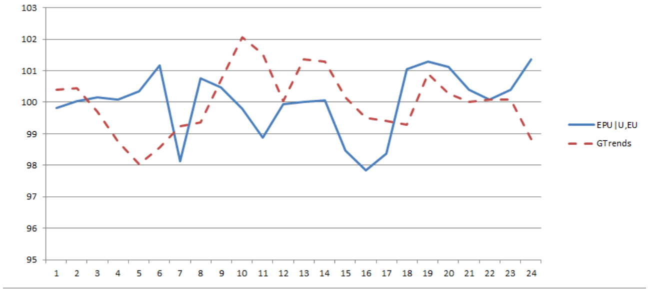
(a) Google trends 2008-2009 vs U



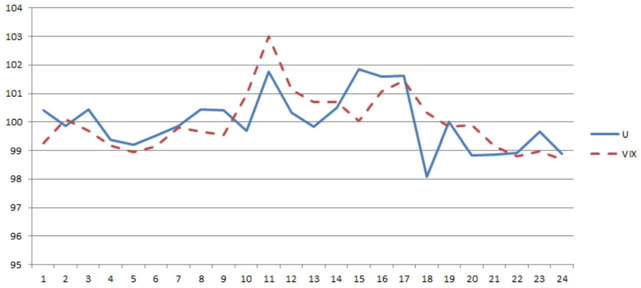
(b) Google trends 2008-2009 vs EU



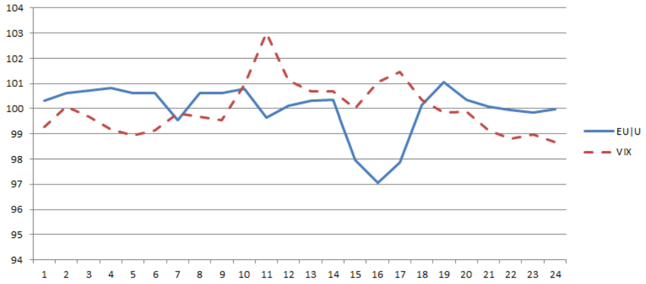
(c) Google trends 2008-2009 vs PU



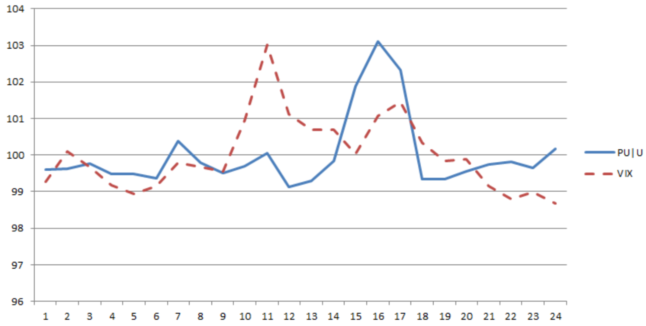
(d) Google trends 2008-2009 vs EPU



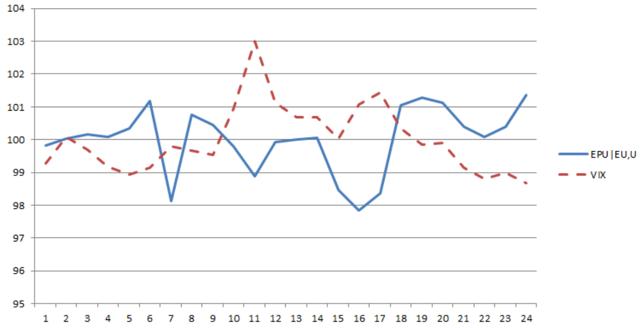
(a) Indian VIX 2008-2009 vs U



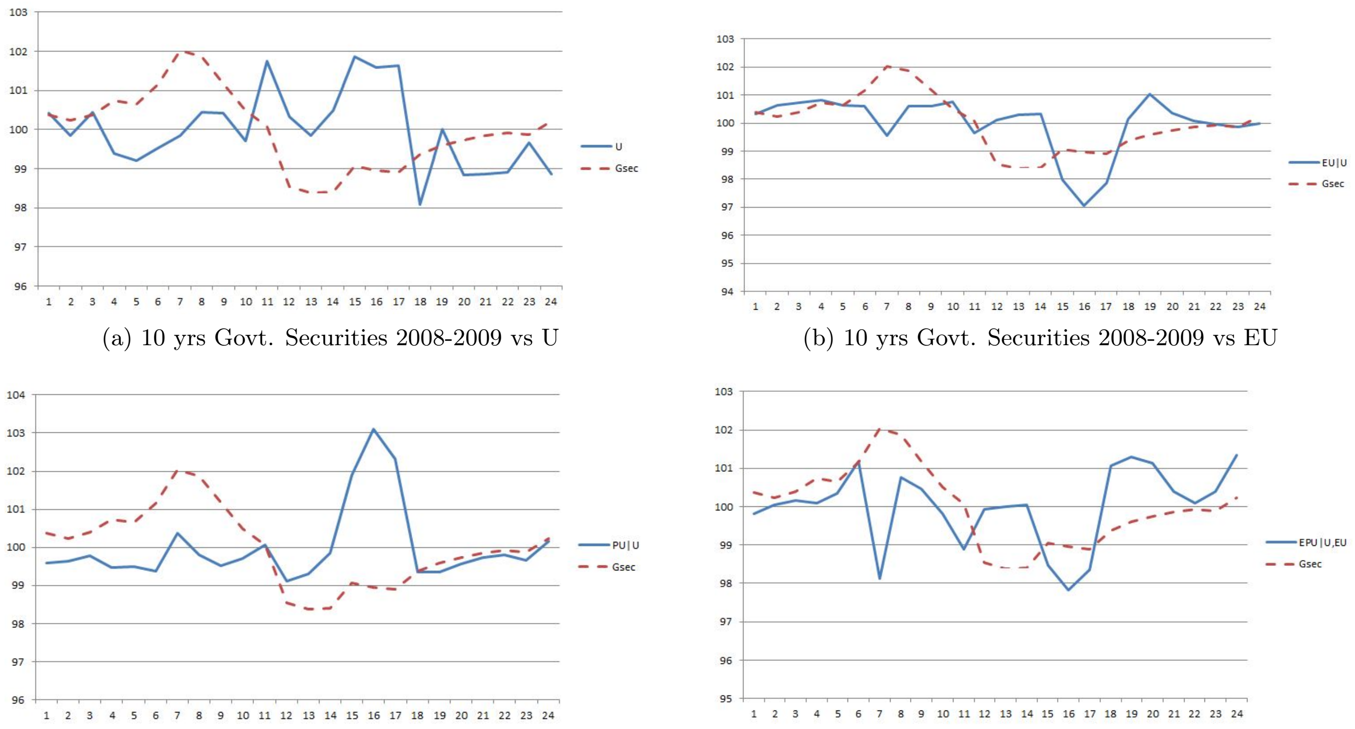
(b) Indian VIX 2008-2009 vs EU



(c) Indian VIX 2008-2009 vs PU



(d) Indian VIX 2008-2009 vs EPU



(a) 10 yrs Govt. Securities 2008-2009 vs U

(b) 10 yrs Govt. Securities 2008-2009 vs EU

(c) 10 yrs Govt. Securities 2008-2009 vs PU

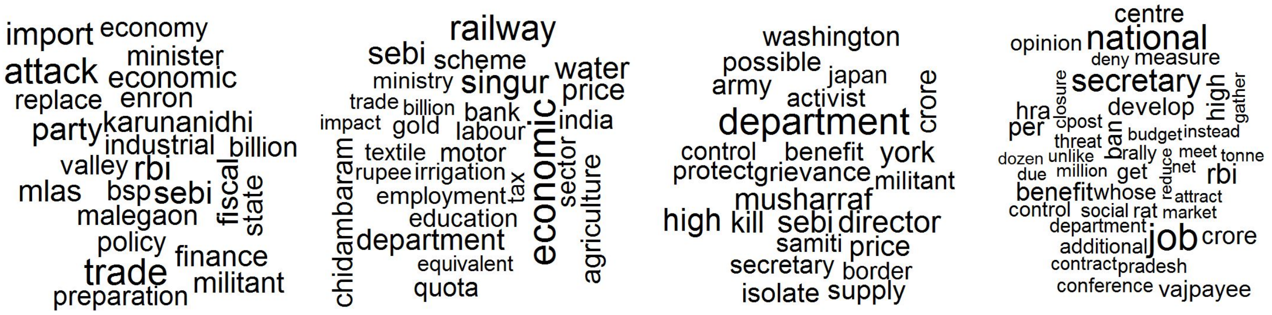
(d) 10 yrs Govt. Securities 2008-2009 vs EPU

Table 4: Correlations with Unconditioned Indexes (2008 & 2009)

	U	EU	PU	EPU
India VIX	0.606	-0.203	0.407	-0.306
Google Trends	0.278	0.252	-0.056	0.171
Govt. Securities	-0.212	0.388	-0.28	0.360

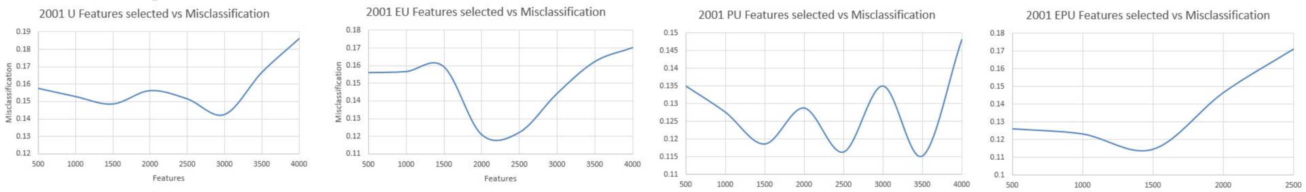
7 Appendix

7.1 Word clouds obtained from SVM model

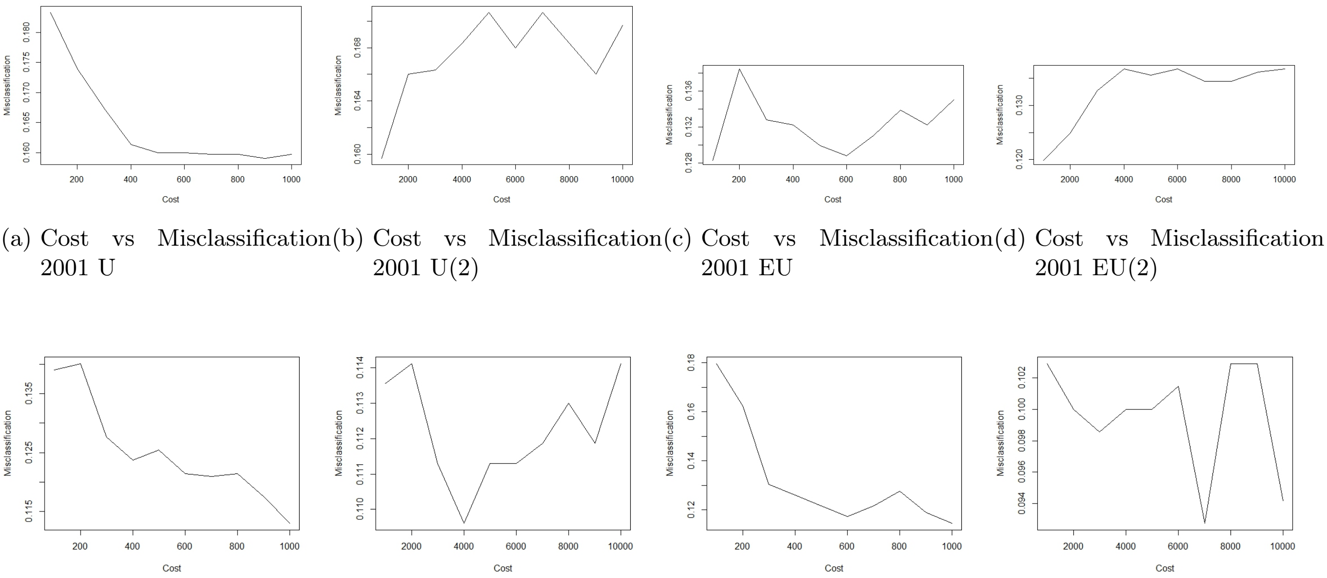


(a) Wordcloud for U (b) Wordcloud for EU (c) Wordcloud for PU (d) Wordcloud for EPU

7.2 Model Selection Graphs

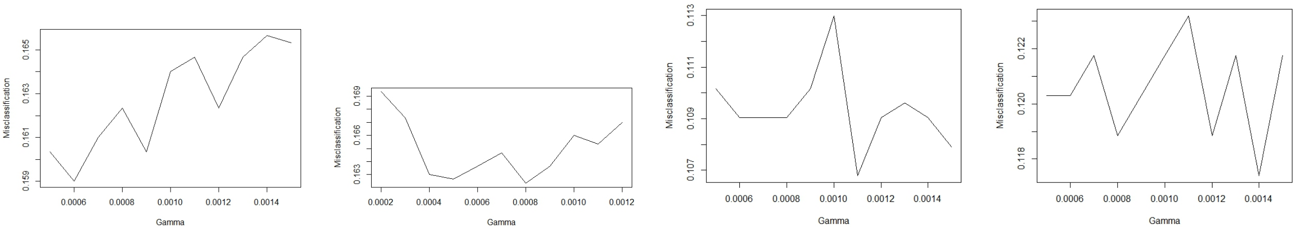


(a) Features selected vs Misclassification 2001 U (b) Features selected vs Misclassification 2001 EU (c) Features selected vs Misclassification 2001 PU (d) Features selected vs Misclassification 2001 EPU

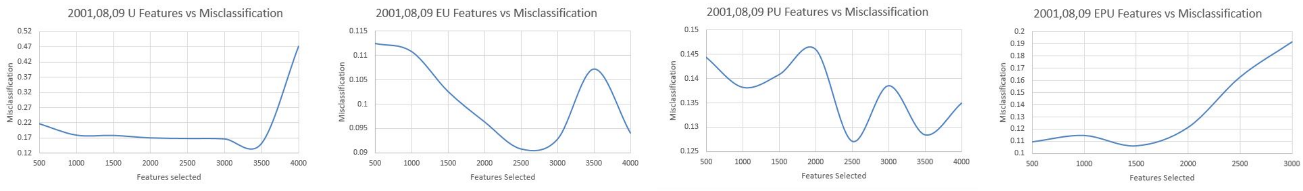


(a) Cost vs Misclassification 2001 U (b) Cost vs Misclassification 2001 U(2) (c) Cost vs Misclassification 2001 EU (d) Cost vs Misclassification 2001 EU(2)

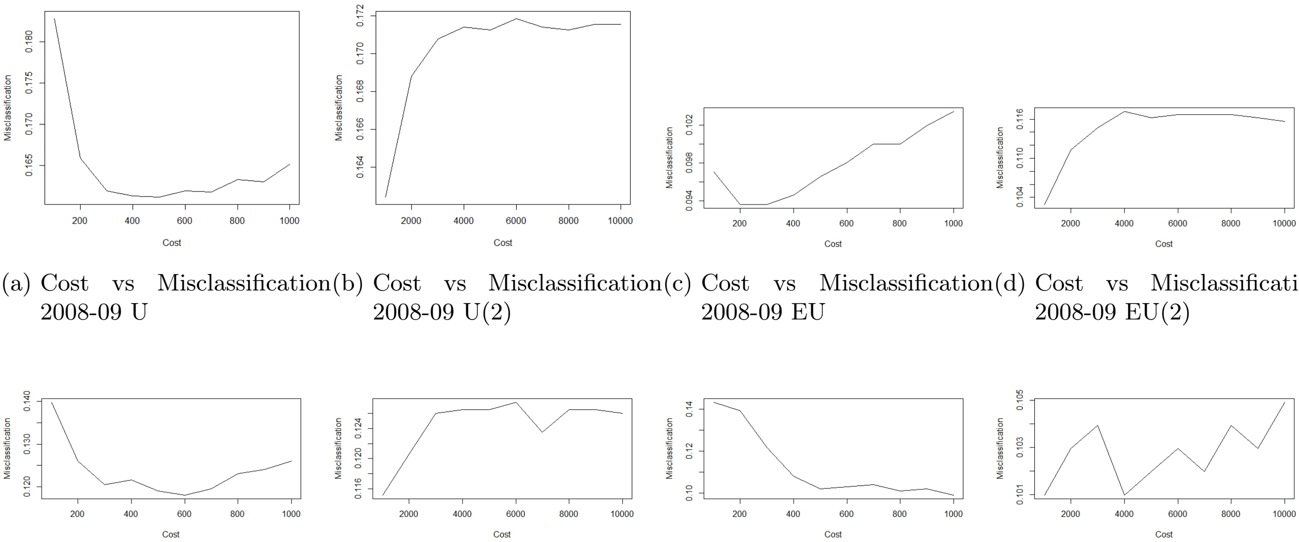
(e) Cost vs Misclassification 2001 PU (f) Cost vs Misclassification 2001 PU(2) (g) Cost vs Misclassification 2001 EPU (h) Cost vs Misclassification 2001 EPU(2)



(a) γ vs Misclassification 2001 U (b) γ vs Misclassification 2001 EU (c) γ vs Misclassification 2001 PU (d) γ vs Misclassification 2001 EPU

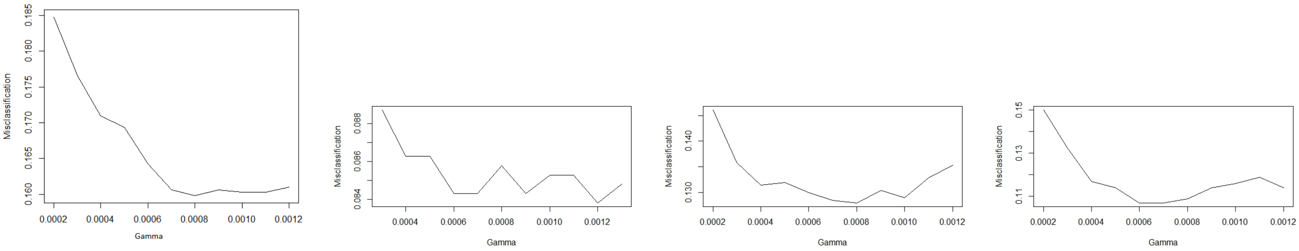


(a) Features selected vs Misclassification 2008-09 U (b) Features selected vs Misclassification 2008-09 EU (c) Features selected vs Misclassification 2008-09 PU (d) Features selected vs Misclassification 2008-09 EPU



(a) Cost vs Misclassification 2008-09 U (b) Cost vs Misclassification 2008-09 U(2) (c) Cost vs Misclassification 2008-09 EU (d) Cost vs Misclassification 2008-09 EU(2)

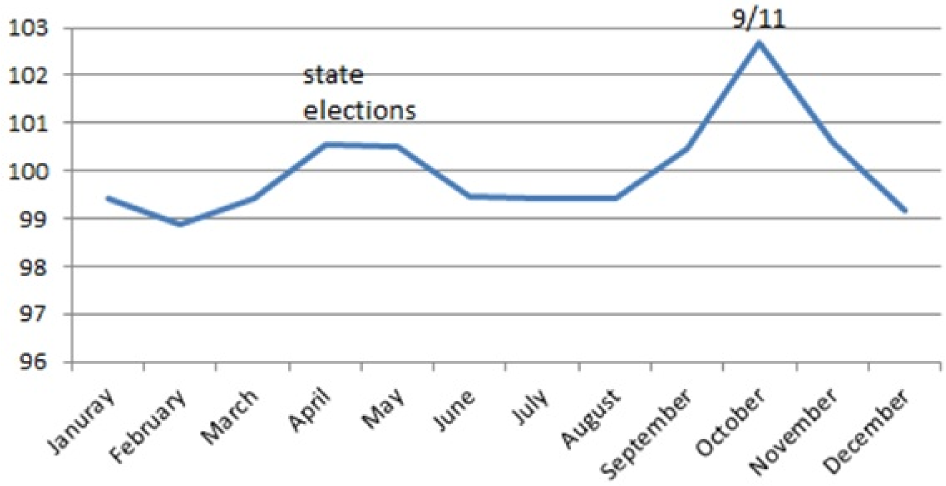
(e) Cost vs Misclassification 2008-09 PU (f) Cost vs Misclassification 2008-09 PU(2) (g) Cost vs Misclassification 2008-09 EPU (h) Cost vs Misclassification 2008-09 EPU(2)



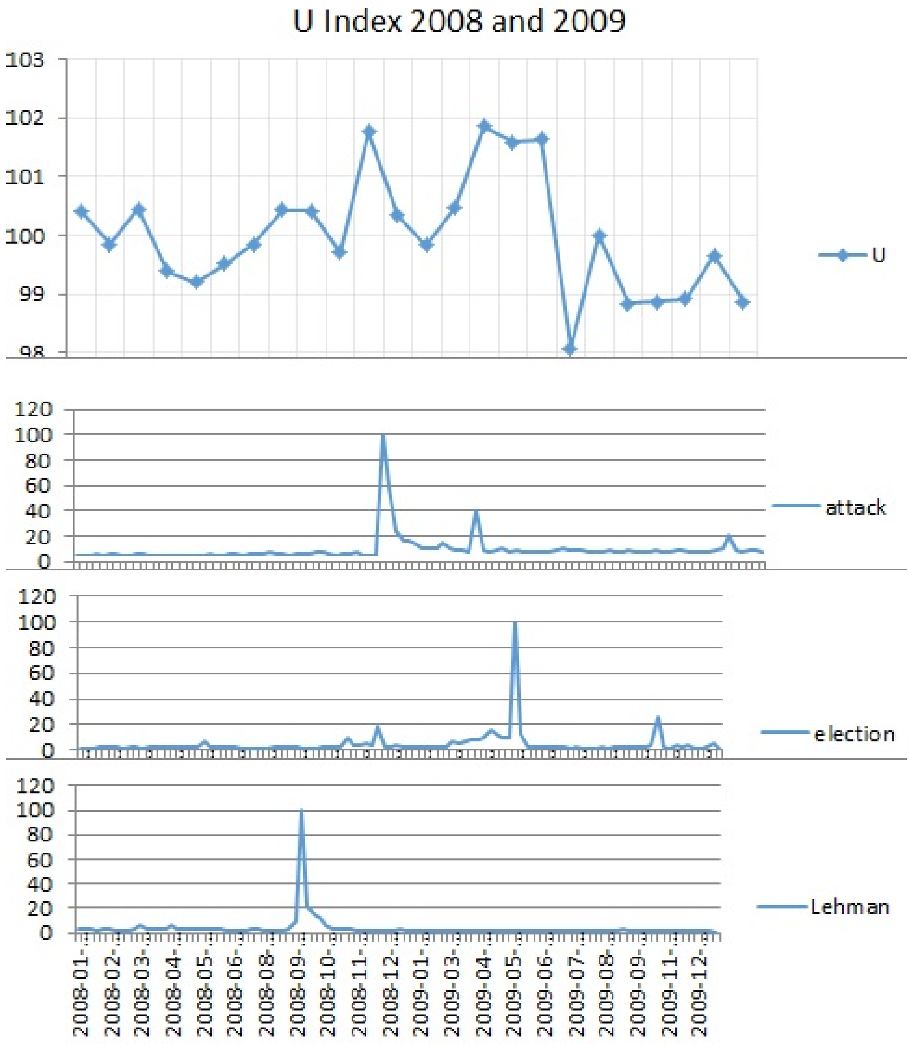
(a) γ vs Misclassification 2008-09 U (b) γ vs Misclassification 2008-09 EU (c) γ vs Misclassification 2008-09 PU (d) γ vs Misclassification 2008-09 EPU

7.3 Other comparisons

Naive U Index 2001



Google trends and U index 2008 and 2009



8 References

1. Aastveit,K.A., Gisle, N. and Sola, S. 2013. “Economic Uncertainty and Effectiveness of Monetary Policy”. working paper. Norges Bank Research

2. Aisen, A. and Veiga, F.J. 2011. “How Does Political Instability Affect Economic Growth?” IMF working paper.

3. Baker, S.R., Bloom, N. and Davis, S.J. , 2013, Measuring Economic Policy Uncertainty, Stanford University working paper.

4. Ajinkya, B.B.,Atiase, R.K. and Gift, M.J. 1990. “Volume of Trading and dispersion in Financial Analysts’ Earnings Forecasts”. The Accounting review.

5. Bhagat, S., Ghosh, P. and Rangan, S.P., 2013, Economic Policy Uncertainty and Economic Growth in India, Indian Institute of Management Bangalore working paper no 407.

6. Choi, H. and Varian, H. 2009. “Predicting the future with Google Trends”

7. Durnev, A., “The Real Effects of Political Uncertainty: Elections and Investment Sensitivity to Stock Prices”, Working Paper, McGill University.

8. Johanssen, B. K. 2014. “When are the effects of fiscal policy uncertainty large”. Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs Federal Reserve Board, Washington, D.C.

9. Tobback, E., Daelemans, W., de Fortuny, E.J., Naudts, H., Martens, D., 2014, Belgian Economic Policy Uncertainty Index : Improvement through text mining, European Central Bank working paper.

10. Zarnowitz, V.1992. “Consensus and Uncertainty in Economic Prediction”.Business Cycles: Theory, History, Indicators, and Forecasting:492-518

11. Audit Bureau of Circulations, 2013, Details of the most circulated publications for the Audit period Jan-June 2013, accessed at http://www.auditbureau.org/news, Press Information Bureau

12. The Registrar of Newspapers for India, Press in India 2013-14, 58th Annual Report, accessed at http://rni.nic.in/pin1314.pdf.

13. Machine Learning - A Probabilistic Perspective, Kevin P. Murphy

14. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”, Thorsten Joachims 1998

15. “Text Categorization and Support Vector Machines”, István Pilászy