Identification and Interpretation of NSWs Using Variational Bayesian Inference in Bengali News Corpus

Chandan Kundu
Research Scholar
West Bengal State University
Barasat, Kolkata, India.
chandankundu2008@gmail.com

ABSTRACT: *The Bayesian model for prediction problems requires setting up the prior/hyper prior structures that go through the process of integration. However, these formulated integrals are not tractable analytically. Moreover, application of Markov Chain Monte Carlo (MCMC) methods to solve these integralsare slow in nature, especially if the parameter space is high dimensional. The key idea behind the Bayesian inference is to marginalize over unknown parameters, rather than make point estimation. This technique avoids severe over-fitting problems and allows direct model comparison.*

In this paper, we presented a model using variational Bayesian inference for identification and interpretation of Non Standard Words in Bengali news corpus. The variational methods extend the practicality of Bayesian inference to complex Bayesian models and "medium sized" data sets. The Variational Bayesian inference aims to approximate posterior distribution by a simpler distribution for which marginalization is tractable.

Keywords: Variational Bayesian Inference, Prior, Posterior, Maximum Likelihood, Expectation Maximization, Non Standard Words, Semiotic Class, Kulback-leibler Divergence

Received: 2 August 2014, Revised 10 September 2014, Accepted 14 September 2014

1. Introduction

Identification and interpretation of Non standard words is a major challenge in the field of information retrieval system. The Real text contains words whose meanings are present in the dictionary. That apart, real text also contains words whose meanings and interpretations are not available in the dictionary. The second category of words is known as Non Standard Word (NSW). NSWs have greater inclination towards ambiguity in terms of their interpretation and pronunciation than ordinary words. In information retrieval system, identification and interpretation of NSWs are important aspect while we are going for further analysisnamely Text to Speech system. In real text, NSWs are represented in diversified forms, for example, digits, words or combination of both. Interpretation of NSWs is much dependent on the context of the NSW [1] because same NSW could be present in different texts representing different meanings (e.g., 2014 may represent year or amount or house number or something else).

Considerable amount of works have been done by different researchers on different languages. However, very insignificant amount of work has been done on internationally popular language Bengali (Bangla). Different researchers adopted different techniques to identify and interpret the NSWs, likely n-gram language model [3], decision tree[3], regular expression[4], Bayesian classification[5]. Comparatively low accuracy levels of certain semiotic classes require further investigations and analysis. Unfortunately only a few researchers used the concept of context in their analysis [6].

In our previous work, we used the concept of context window and context identification array and we achieved significantly good accuracy levels compare to other models [1]. In this paper, we extend the concept of Bayesian classification technique to Variational Bayesian inference and we get very significantly good accuracy values.

Maximum likelihood (ML) estimation is one of the most popular technologies used in modern classification problem. The expectation maximization (EM) is an iterative algorithm used for estimating ML. Since the formal introduction in 1977 by Dempster et al. [7], the EM is gaining popularity for EM ML estimation. Now the EM algorithm has become an important tool used in a wide range of applications, such as classification, recovery and segmentation of images and videos, image modeling, carrier frequency synchronization and channel estimation in communication and speech recognition.

The concept behind the EM algorithm is very intuitive and natural. EM-like algorithms existed in the statistical literature even before [7], however such algorithms were actually EM algorithms in special contexts. The first known such reference dates back to 1886, when Newcomb considered the estimation of the parameters of a mixture of two univariate normal [8]. However, it was in [7] where such ideas were synthesized and the general formulation of the EM algorithm was established.A good survey on the history of the EM algorithm before [7] can be found in [9]. However EM demands certain requirements that limit the applicability to the complex problem.

In general, we can define inference as the process of obtaining a conclusion based on the information available, which includes observed data as a subset. From this, Bayesian inference can be defined as a process of inference using Bays' theorem in which information is used to newly infer the plausibility of a hypothesis. This process produces information that adds to organizational knowledge.

Information about a hypothesis beyond the observable empirical data about that hypothesis is included in the inference. In this view, probability quantifies our state of knowledge and represents the plausibility of an event, where '*plausibility*' implies apparent validity. In other words, Bayesian inference is a mechanism to encode information, where the encoding metric is a value (or an absolute scale from 0 to 1) known as probability.

The use of Bayesian inference methods uses all of the available information and leads to better parameter estimates and to better decision. These aspects are important since decision makers are often asked to make inference using sparse data.

Bayesian inference has certain characteristics namely (i) BI generates information in terms of probabilities related to a hypothesis. *BI = information, where information = models + data + other information* (ii) probability is a measure of the degree of plausibility of a hypothesis (iii) since probability is subjective, for any hypothesis there is no true value for its associated probability.

Inference, or model evaluation, is the process of updating probabilities of outcomes based upon the relationships in the model and the evidence known about the situation at hand. When a Bayesian model is actually used, the end user applies evidence about recent events or observations. This information is applied to the model by "*instantiating*" or "*clamping*" a variable to a state that is consistent with the observation. Then the mathematical mechanics are performed to update the probabilities of all the other variables that are connected to the variable representing the new evidence.After inference, the updated probabilities reflect the new levels of belief in (or probabilities of) all possible outcomes coded in the model. These beliefs are mediated by the original assessment of belief performed by the author of the model.

Posterior distribution is calculated with the help of likelihood and prior distribution according to the Bayes formula

$$p(i \mid j) = \frac{p(i)\, p(j \mid i)}{p(j)} \qquad (1)$$

Where $p(i \mid j)$ is the posterior distribution and $p(j \mid i)$ is the likelihood. This equation of logical inference is known as Bayes'

theorem. If we dissect (1), we will see there are four parts, as listed in the figure 1:

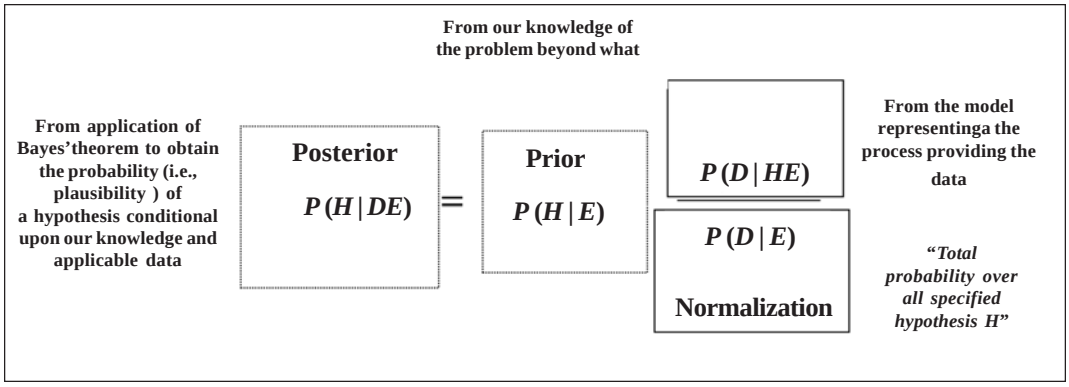

Figure 1. The equation for Bayes' theorem dissected into four parts

Where D = The Data

H = Our Hypothesis

E = General information known prior to having updated information (or data) specific to problem at hand in other words, our Experience.

The denominator of Bayes' theorem is sometimes denoted $f(x)$, and is called the marginal or unconditional distribution of x. The range of integration is over all possible values. In cases where X is a discrete random variable (e.g., number of events in some period of time), $f(x)$ is the probability of seeing x events, unconditional. In this context, which will become useful for model validation, $f(x)$ will be referred to as the predictive distribution for X.

Traditionally the likelihood function is most often binomial, poison or exponential[1]. Note that the symbol "|" represents a conditionality, which in the case of the likelihood function is described as "*the probability we see the observed data given the parameter takes on a certain value*." Priors can be classified broadly as either informative or non-informative. Informative priors, as the name suggests, contain substantive information about the possible values of the unknown parameter. Non-informative priors, on the other hand, are intended to let the data dominate the posterior distribution; thus, they contain little substantive information about the parameter of interest.

Most of the researchers used Markov Chain Monte Carlo (MCMC) algorithms to estimate the posterior distribution. But for large dimensionality involving a complicated covariance matrix, MCMC is not a good choice for estimation and it incurs high computational cost [11]. Therefore, researchers thought for an alternative way and in 1995D. Mackay found a new algorithm known as Variational Bayesian (VB) [21] which is suitable for large dimensional spaces and is very cost effective [11]. Besides, the rate of convergence in VB is better than MCMC.

In this paper, we have aimed to highlight on Variational Bayesian technology for statistical inference that overcomes the shortcomings of the EM algorithm. Bayesian inference [10] based on the variational approximation has been used extensively by the machine learning community since the mid-1990s when it was first introduced. Now, VB is widely accepted and widely used algorithm in different fields of computer sciences, signal processing and so many. The main thought of VB is to approximate the joint posterior of the unknown by a separable density.

2. Bayesian Inference Basic

Assume that $x = \{x_1, x_2, \ldots, x_n\}$ are the observation and θ the unknown parameters of a model that generated x. It is worthy to mention that the term estimation is used to refer to the parameters and inference refers to random variables. The term estimation

[1]when dealing with repair or recovery times, thelikelihood function may be lognormal, weibull or gamma. If the observed variable is unavailable, the likelihood may be beta distribution

is used to refer to the parameters and inference refers to random variables. The term estimation refers to the calculated approximation of the value of parameters from incomplete, uncertain and noisy data. In contrast, the term inference will be used to imply Bayesian inference and refers to the process in which prior evidence and observation are used to infer the posterior probability $p(\theta \mid x)$ of the random variables θ given the observation x.

2.1. Parameter Estimation

One of the most popular approaches for parameter estimation is ML. The ML approach gives the estimation as

$$\theta ML = \frac{argmax\, p(x \mid \theta)}{\Theta} \tag{2}$$

where $p(x|\theta)$ refers to the probabilistic relationship between the observations and the parameters based on the assumed model that produces the observation x. By Bayes' rule, the posterior over parameters θ given the observation x is given by

$$p(\theta \mid x) = \frac{p(\theta)\, p(x \mid \theta)}{p(x)} \tag{3}$$

The second term in the numerator is the marginal likelihood or evidence [14]. In many cases direct assessment of the likelihood function $p(x \mid \theta)$ is complex and is either difficult or impossible to compute it directly or optimize it. In such cases the computation of likelihood function is greatly affected by the introduction of the hidden variables z_i (i =1...n).These random variables act as links that connect the observations to the unknown parameters via Bayes' law.

The selection of hidden variables is problem specific. However, as their name suggests, these variables are not observed but they supply enough information about the observations. Therefore the conditional probability $p(z \mid x)$ is easy to compute. Apart from this role, hidden variables play another important role in statistical modeling. They are an important part of the probabilistic mechanism that is assumed to have generated the observations and can be described very succinctly by a graph that is termed '*graphical model*'.

Once hidden variables and a prior probability for them $p(z \mid \theta)$ have been introduced, one can get the likelihood or the marginal likelihood as it is called at times by integrating out (marginalization) the hidden variables according to

$$p(x \mid \theta) = \int p(x, z \mid \theta)\ dz = \int p(x \mid z, \theta) \int p(z \mid \theta)\ dz \tag{4}$$

Now for a new *datum x'*, the predictive density will be

$$p(x' \mid x, \theta) = \int p(z \mid x, \theta)\ p(x' \mid z, x, \theta)\, dz \tag{5}$$

This can be written as

$$p(x' \mid x, \theta) = \int p(z \mid x, \theta)\ p(x' \mid z, \theta)\, dz \tag{6}$$

Now we are interested to find the likelihood function and posterior of the hidden variables according to

$$p(z \mid x, \theta) = \frac{p(x \mid z, \theta)\ p(z \mid \theta)}{p(x \mid \theta)} \tag{7}$$

If we consider x' is conditionally independent of x given θ, the posterior distribution of hidden variables z associated with new observation x'

$$p(z \mid x', x, \theta) = \frac{p(x' \mid z, x, \theta)\ p(z \mid \theta)}{p(x' \mid \theta)} \tag{8}$$

Once we calculate the posterior, it is possible to calculate the inference for the hidden variables. Despite the simplicity of the above formulation, generally the integrals in (4) and (5) are either impossible or very difficult to compute in closed form. Thus, the main effort in Bayesian inference is concentrated on techniques that allow us to bypass or approximately evaluate this integral. There are two main categories to evaluate this integral. The first category is Monte Carlo technique [13][14] that concentrates on numerical sampling methods and the second category belongs to deterministic approximation method. In this paper we have concentrated only on the deterministic approximation method. Moreover, maximum a posteriori (MAP) inference,

[1]when dealing with repair or recovery times, thelikelihood

which is an extension of the ML technique, can be considered as a very crude Bayesian approximation because, strictly speaking, MAP estimation deals only with è random variables. However, in ML, Bayesian inference is used for hidden variables.

As we have discussed earlier, EM algorithm is Bayesian inference methodology that considers the posterior $p\ (z \mid x, \theta)$ and iteratively maximized the likelihood function without explicitly computing it. A major drawback of this technology is that in many cases posterior is not available. However, current advancement of Bayesian inference allows us to bypass this difficulty by approximating the posterior. This current approximating technique is known as '*Variational Bayesian approximation*'.

2.1.1 EMAlgorithm

The goal of the EM algorithm is to find the maximum likelihood solution for a model consisting of parameters Θ, given observed (incomplete) data *x* and latent or hidden variables *z*.

Calculation of log-likelihood [15][16] is straightforward and it can be written as

$$L\ (\Theta) = In\ p\ (x \mid \Theta) = \sum_{i=1}^{n} In\ p\ (x_i \mid \Theta) = \sum_{i=1}^{n} \int dz_i\ p\ (z_i, x_i \mid \Theta)$$

After simplification, we get

$$L\ (\Theta) = F\ (q, \theta) + KL\ (q \,||\, p) \tag{9}$$

where

$$F\ (q, \theta) = \int q\ (z)\ In\ \left(\frac{p\ (x, z \mid \theta)}{q\ (z)}\right)\ dz \tag{10}$$

and

$$KL\ (q \,||\, p) = -\int q\ (z)\ In\ \left(\frac{p\ (z|x, \theta)}{q\ (z)}\right)\ dz \tag{11}$$

where $q\ (z)$ is any probabilistic density function over the hidden variables gives rise to a lower bound on $\mathcal{L}$. $KL\ (q||p)$ is the Kullback-leibler divergence between $p\ (z \mid x, \theta)$ and $q\ (z)$. Since $KL\ (q \,||\, p) >= 0$, we can say $\ln p\ (x| \theta) >= F\ (q, \theta)$. Hence $F\ (q, \theta)$ is a lower bound on $L\ (\Theta)$ i.e., the log-likelihood. Equality holds only when $KL\ (q \,||\, p) = 0$, which implies $p\ (z \mid x,\theta) = q\ (z)$. The EM algorithm and some recent advances in deterministic approximations for Bayesian inference can be viewed in the light of the decomposition in (9) as the maximization of the lower bound $F\ (q, \theta)$ with respect to the density q and the parameters θ.

In general EM algorithm is a two step algorithm that maximizes the lower bound $F\ (q, \theta)$ and hence the log-likelihood. Let the current value of the parameter is θ^{old}. In the E-step the lower bound $F\ (q, \theta)$ is maximized with respect to $q\ (z)$. The maximum value of $F\ (q, \theta)$ is achieved when $KL\ (q \,||\, p) = 0$ i.e., when $q\ (z) = p\ (z \mid x, \theta^{old})$. In this condition the lower bound is equal to the log-likelihood. In the consecutive M-step, $q\ (z)$ is kept fixed and lower bound $F\ (q, \theta)$ is maximized with respect to θ to give new value θ^{new}.The effect of M-step is to increase the lower bound and as a result, the corresponding log-likelihood will also increase. As $q(z)$was calculated using θ^{old} and is kept fixed in the M-step, it will not be equal to the new posterior $p\ (z|x, \theta^{new})$ and hence KL distance will not be zero. So we can say that the increase in the log-likelihood is greater than the increase in lower bound. Now, we substitute $q(z) = p\ (z|x, \theta^{old})$ into the lower bound and expanding (10) we get

$$F\ (q, \theta) \int p\ (z \mid x, \theta^{old})\ in\ p\ (x, z \mid \theta)\ dz\ \ - \int p\ (z \mid x, \theta^{old})\ in\ p\ (z| x, \theta^{old})\ dz\ = Q\ (\theta, \theta^{old}) + Constant \tag{12}$$

Here the constant is simply the entropy of $p\ (z \mid x, \theta^{old})$ that does not depend on θ. The function

$$Q\ (\theta, \theta^{old}) = \int p\ (z \mid x, \theta^{old})\ in\ p\ (x, z \mid \theta) = \langle\ in\ p\ (x, z \mid \theta)\ \rangle_{p\ (z \mid x, \theta^{old})} \tag{13}$$

gives the expectation of log-likelihood of the complete data (observation + hidden variables) which is maximized in the M-step. The EM algorithm can be summarized as

$$\textbf{E - step:}\ \text{compute}\ p\ (z \mid x, \theta^{old}) \tag{14}$$

$$\textbf{M - step:}\ \text{compute}\ \theta^{new} = \underset{\theta}{argmax}\ Q\ (\theta, \theta^{old}) \tag{15}$$

It is interesting to note that the EM algorithm requires that $p\ (z|x, \theta)$ is explicitly known, or at least we should be able to compute

the conditional expectation of its sufficient statistics$\langle in\, p(x, z \mid \theta) \rangle_{p(z \mid x, q^{old})}$. In other words, it required to know the conditional pdf of the hidden variables given the observations in order to use the EM algorithm. While $p(z|x, \theta)$ in many interesting problems this is not possible and thus the EM algorithm is not applicable.

2.1.2 Variational Bayesian Inference

Given a training corpus $\boldsymbol{x}$, computation of the posterior probability distribution over parameters Θ is too complicated. The Variational EM algorithm bypasses the requirement of exactly knowing posterior with respect to hidden variables by assuming an simpler distribution $q(z)$ in the decomposition of (9). q is called the Variational approximation to the posterior.

In the E-step $F(q, \theta)$ is maximized with respect to $q(z)$ keeping è fixed. In the process of maximization, a particular form of $q(z)$ must be assumed. In certain cases it is possible to assume knowledge of the form of $q(z|\omega)$, where ù is a set of parameters. Therefore the lower bound $F(\omega, \theta)$ becomes a function of these parameters and is maximized with respect to ù in the E-step and with respect to è in the M-step, see for example [15]. The generalized algorithm of VB is given below:

Step 1 : Initialize all approximate posteriors by setting them to theirpriors.

Step 2 : while $(|\Theta_i^{new} - \Theta_i^{old}|) > 0.0001)$
evaluate $q^{new}(z)$to maximize$F(q, \Theta^{old})$
find $\Theta_i^{old} = \arg\max_{\Theta} F(q^{new}, \Theta_i)$

Figure 2. Algorithm for Variational Bayesian Inference

The general form of the lower bound $F(q, \theta)$ is a functional in terms of q. A mapping that takes as input a function $q(z)$ and returns as output the value of the functional. This leads naturally to concept of the functional derivative, which in analogy to the function derivative, gives the functional changes to the input function. This area of mathematics is called calculus of variations [17] and has been applied to many areas of mathematics, physical science and engineering, for example, fluid mechanics, heat transfer, control theory, machine learning, classification and so many.

Interesting fact regarding variational theory is that there are no approximations in it but variational method can be used to find approximate solutions in Bayesian inference problem. This can be achieved by assuming that the function, over which optimization is performed, have specific forms, for example, we can assume only quadratic functions or functions that are linear combinations of fixed basis functions. For Bayesian inference, a particular form that has been used with great success is the factorized one [18][19]. The idea for this factorized approximation steams from theoretical physics where it is called mean field theory [20].

According to this approximation, the hidden variables z are assumed to be portioned into M partitions z_i with $i = 1, \ldots, M$. Also it is assumed that $q(z)$ factorizes with respect to these partitions as

$$q(z) = \prod_{i=1}^{M} q_i(z_i) \tag{16}$$

Thus, we require to find the $q(z)$ of the form of (16) that maximizes the lower bound $F(q, \theta)$. Using (16) and denoting for simplicity $q_j(z_j) = q_j$ we have

$$F(q, \theta) = \int \prod_i q_i \; In\, p(x, z \mid \theta) - \sum_i In\, q_i \; dz$$

$$= \int \prod_i q_i \, In\, p(x, z \mid \theta) \prod_i dz_i - \sum_i \int \prod_j q_j \, In\, q_i \, dz_i$$

$$= \int q_j \, In\, p(x, z \mid \theta) \prod_{i \neq j} (q_i \, dz_i) \; dz_j - \int q_j \, In\, q_j \, dz_j$$

$$- \sum_{i \neq j} \int q_i \, In\, q_i \, dz_i$$

$$= \int q_j \, In\, \tilde{p}(x, z_j \mid \theta) \, dz_i - \int q_j \, In\, q_j \, dz_j$$

$$- \sum_{i \neq j} \int q_i \, In\, q_i \, dz_i$$

$$= - KL\,(q_j \,||\, \tilde{p}\,) - \sum_{i \neq j} \int q_i \, In\, q_i \, dz \qquad (17)$$

where

$$In\, \tilde{p}\,(x, z_j \,|\, \theta\,) = \langle\, In\, p\,(x, z \,|\, \theta\,)\,\rangle_{\,i \neq j}$$

$$= \int In\, p\,(x, z \,|\, \theta\,) \prod_{i \neq j} (\, q_i \, dz_i\,)$$

Clearly the bound in (17) is maximized when the Kullback-Leibler distance becomes zero, which is the case for, $q_j = \tilde{p}\,(x, z_j \,|\, q\,)$ in other words the expression for the optimal distribution $q_j^{\,*}\,(z_j)$ is

$$In\, q_j^{\,*}\,(z_j) = \langle\, In\, p\,(x, z \,|\, \theta\,)\,\rangle_{\,i \neq j} + Constant \qquad (18)$$

The additive constant in (18) can be obtained through normalization, thus we have

$$q_j^{\,*}\,(z_j) = \frac{exp\,(\,\langle\, In\, p\,(x, z \,|\, \theta\,)\,\rangle_{\,i \neq j}\,)}{\int exp\,(\,\langle\, In\, p\,(x, z \,|\, \theta\,)\,\rangle_{\,i \neq j}\,)\, dz_j} \qquad (19)$$

The above equation for j =1, 2, ……, M are a set of consistency conditions for the maximum of the lower bound subject to the factorization of (16). They do not provide an explicit solution since they depend on the other factors $q_j(z_j)$ for $i \neq j$. Therefore, a consistent solution is found by cycling through these factors and replacing each in turn with the revised estimate.

The summarization of the Variational EM algorithm is given by the following two steps:

Variational E-step: Evaluate $q^{new}(z)$ to maximize $F(q, \theta^{old})$ solving the system of (19).

Variational M-step: Find $\theta^{new} = arg\ max_{\theta}\, F\,(q^{new}, \theta)$.

Now, it is worthy to mention that in certain cases a Bayesian model can contain only hidden variables and no parameters. In such cases the Variational EM algorithm has only an E-step in which $q\,(z)$ is obtained using (19). This function $q\,(z)$ constitutes an approximation to $p\,(z \,|\, x)$ that can be used for inference of the hidden variables.

3. Experiments and Result

We have implemented our proposed technique to classify NSWs in Bengali news corpus. We have divided the whole task of classification into four major steps, namely, primary classification using regular expression, feature vector generation, final classification using Variational Bayesian inference and interpretation of NSWs based on specific semiotic class. The pictorial representation of the proposed technique is given in the appendix A. The whole process is implementedusing Python3 programming language.

Preparation of the databases from news corpus is a very important as well as time consuming task before the start of actual classification process. The online version of the famous Bengali news paper 'AnandabazarPatrika'[1] has been used for source document of the databases (training data and test data). The time spans required for the news corpus were approximately 23 days and 12 days, for training data set and test data set respectively.Initially the raw data weredownloaded from the web site and stored in the plain text format. In the preprocessing steps, we removed all the unwanted characters and images.The advertisements on the web page were also separated from the content and removed from the plain text. The format of the text was stored in UTF-8 standard. The size of the training and test data sets are 2.03MB and 803KB respectively.

The primary classification using regular expressionis performed on our training database[1] and details of classification had been presented in our early work [1]. In the primary classification step, we have generated database D containing examples (or sentences) of different semiotic classes (e.g., money, time, telephone number, year etc.). From the generated data base D, we create word features which is a list of every distinct words present in D[22][23]. To train a classifier we require to identify what features are relevant [22][23]. For that, we have generated a feature vector indicating what words are contained in the input D passed. The typical structure of the feature vector is as follows:

feature column *j*

...
examples *i* 0 1 0 1 0 0 0 0 1....
1 0 0 0 1 1 0 0 0....
...

Each row represents one example (or, one sentence containing NSW), and each column represents one feature, where '1'denotes the existence of the feature in this context,and 0 denotes the nonexistence.Note that columns should be separated by spaces.

In the final classification step, the feature vector becomes the input. In the current section we are using the following notational conventions:

Symbol	Meaning
w	a NSW
s_1.....s_k	semiotic class
c_1.....c_i	context of w in database D
f_1.....f_j	words used as contextual features for identification

In contrast to Bayes classifier of [24], parameter estimation in unsupervised learning technique is not based on the given label (semiotic class) set. Instead, theidentities$s$ (i = 1...k, number of semiotic class) of the semiotic classes are unknown. we start with a random initialization of the parameters p (s_k) and p (f_j |s_k). The p (f_j |s_k) are then re-estimated by the Variational Bayesian algorithm. After the random initialization, we compute for each context c_i of w the probability P (c_i | s_k) that it was generated by semiotic class s_k. We can use this preliminary categorization of the contexts as our training data and then re-estimate the parameters p (f_j | s_k) soas to maximize the likelihood of the data given the model.

The Variational Bayesian algorithmis guaranteed to increase the log likelihood of the model. Therefore, the stopping criterion for the algorithm is to stop when the likelihood is no longer increasing significantly i.e., ($|s_k^{new} - s_k^{old}| < 0.0001$) (where 0.0001 is the threshold value).

Semiotic Class	Using context window and Context identific-ation Array[1] (on Bengali news corpora)	Variatio-nal Bayesian inference(on Bengali news corpora)	Variatio-nal Bayesian inference (on English news corpora[5])
Date & Month	95.95	98.23	98.13
Money	100	100	100
Telephone No.	100	100	100
Year	97.54	98.24	99.1
Time	94.05	96.81	100
URL	100	100	100
Percentage	100	100	100
Quantity	100	100	98.77
Float	100	100	100

Table 1. Accuracy values for different semiotic classes

[1] Available at http://www.anandabazar.com/

[2] Available at http://eiilm.co.in/Training_Data_Set.txt

The proposed system is thus trained to build up the knowledge base that would be used later to classify test examples[4].

We have carried out the experiments on English news corpus (electronic version of the 'The Times of India') also to justify the efficacy of our proposed model. For that, we have created separate regular expressions (re), training data sets and test data set. In case of English news corpus, instead of creation of explicit databases, we have implemented the primary classification steps directly on the web pages using different python functions.

The following table gives the details of classification accuracies (in percentage) for different semiotic classes.

Semiotic Class	Using context window and Context identification Array[1] (on Bengali	Variational Bayesian inference (on Bengali news corpora)	Variational Bayesian inference (on English news corpora[5])

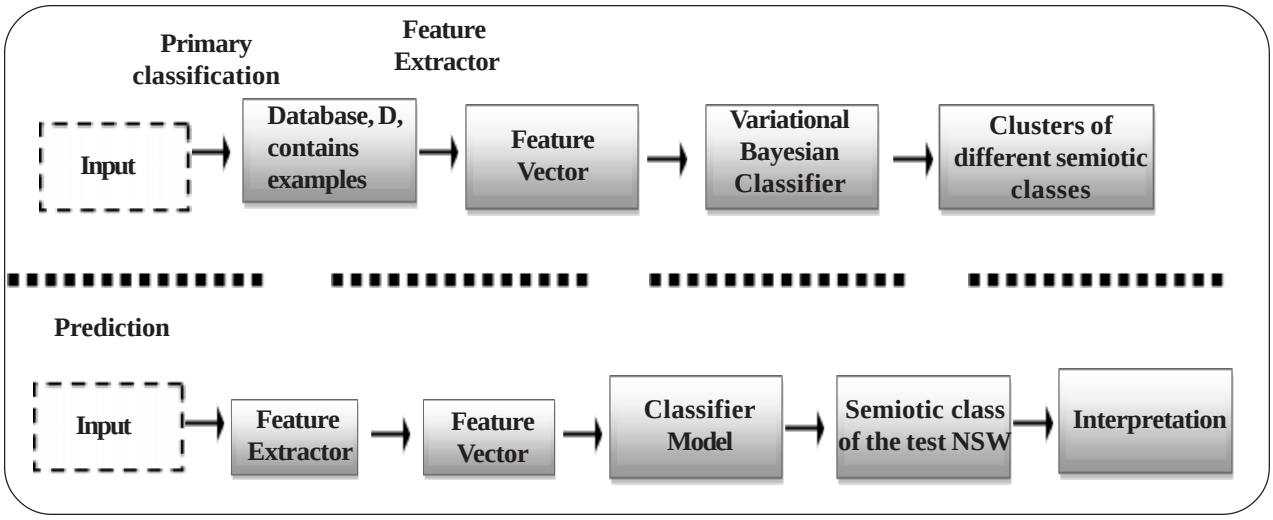

Figure 3. Flow chart of the proposed model

4. Conclusion and Future Work

In this paper we have presented a model for classification and interpretation of NSWs in Bengali news corpus. In particular, this model is applicable for any language. The proposed technique gives satisfactory results over previous models [1][2]. Inspite of getting good accuracy levels, we could not get 100% accuracy values for some semiotic classes (e.g., 'year', 'time', etc.) because the degree of inflation is quite high in Bengali news corpus.

While we are going to build models using either EM theory or Variational Bayesian theory, initial guesses of the different parameters play a vital roleto measure the effectiveness of the model in terms of time complexity. If the initialization values are close to the actual values (which we get after a number of iterations), the time complexity can be reduced to a greater extent. Future researchers have scope to work on this aspect.

Still the research opens up a ample of facets of further research by different approaches and methodologies like maximum entropy, SVM, Gaussian process prior model, Variational Gaussian process model, etc. Once effectiveness of these models will be assessed, the optimal technique can be addressed for identification and interpretation of NSWs in Bengali news corpus.

[1] Available at http://eiilm.co.in/Test_Data_Set.txt

[2] Available at http://epaperbeta.timesofindia.com/

References

[1] Kundu, C., Das, R. K., Sengupta, K. (2013). Implementation of Context Window and Context Identification Array for Identification and Interpretation of Non Standard Word in Bengali News Corpus, *International Journal of Computational Linguistics Research*, 4 (4), p. 159-171.

[2]Alam, F., Habib, S. M. M., Khan, M. (2009). Text Normalization System for Bangla, *In*: Proc. Conference on Language and Technology (CLT09), NUCES, Lahore, Pakistan, January, p. 22-24.

[3] Sproat, R., Black, A.W., Chen, S., Kumar, S., Osetendorfk, M., Richards, R.(2001).Normalization of non-standard words, Computer Speech and Language, p. 287–333.

[4] Xydas, G., Karberis, G., Kouroupetroglou, G. (2004).Text Normalization for pronunciation of Non-Standard Words in an Inflected Language, *In*: proceedings of the 3rd Hellenic conference on Artificial Intelligence (SETN04), Samos, Greece.

[5] Golding, A. (1994). A Bayesian hybrid method for context-sensitive spelling correction, *In*: proceedings of the 3rd workshop on very large corpora, p. 39-53.

[6] Raj, A. A., Sarkar, T., Pammi, S. C., Yuvaraj, S., Bansal, M., Prahallad, K., Black, A. W. (2008). Text Processing for Text-to-Speech Systems in Indian Languages, ISCA SSW6, Bonn, Germany, p. 188-193.

[7] Dempster, A., Laird, N., Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statis*. Soc. A, 39 (1) p. 1–38.

[8] Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result, *Amer. J. Math.*, 8, p. 343–366.

[9] McLachlan, G., Krishnan, T. (1997). The EM Algorithm and Extensions, New York: Wiley.

[10] Stigler, S. M.(1982). Thomas Bayes's inference, *J. Roy. Statis. Soc. A*, 145, p. 250–258.

[11] Robert, C., Cassela, G., (1999). Monte Carlo Statistical Methods, NewYork: Springer-Verlag.

[12] Andrieu, C., Freitas, N. D., Doucet, A., Jordan, M. (2003). An introduction to MCMC for machine learning, *Mach. Learn*., 50, (1) 5–43.

[13] Just, W. (2012). A Survey of Research Applying Bayesian Methods to Natural Language Processing, CS527 Fall.

[14]Beal, M. J. (2003).Variational algorithms for approximate Bayesian inference, Ph.D. thesis, The Gatsby Computational Neuroscience Unit, University of London.

[15] Bishop, C. (2006).Pattern Recognition and Machine Learning, New Yoek:Springer- Verlag.

[16] Neal, R. M., Hinton, G. E. (1998). A view of the EM algorithm that justifiesincremental, sparse and other variants in Learning in Graphical Models, Jordan, M. I., Ed. Cambridge, MA: MIT Press, p. 355–368.

[17] Weinstock, R. (1974). Calculus of Variations, New York: Dover.

[18] Jaakkola, T. S. (1997). Variational methods for inference and learning in graphicalmodels, Ph.D. dissertation, Dept. Elect. Eng. Comp. Sci., MIT.

[19] Jordan, M., Gharamani, Z., Jaakkola, T., Saul, L.(1998). An introduction to variational methods for graphical models in Learning in Graphical Models, Jordan, M., Ed., Cambridge, MA: MIT Press, p. 105–162.

[20] Parisi, G., (1988).Statistical Field Theory, Reading, MA: Addison-Wesley.

[21] Mackay, D. J. C. (1995). Probable networks and plausible predictions- a review of practical Bayesian methodsfor supervised neural networks, Network: Computation in neural Systems, p. 469-505.

[22] Ravikiran. (2012). How to build a twitter sentiment analyzer?, http://ravikiranj.net/drupal/ 201205/code/machine-learning/how-build-twitter-sentiment-analyzer.

[23] Luce, L. (2012). Twitter sentiment analysis using python and NLTK, http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk.

[24] Gale., William, A., Church, K. W., Yarwsky, D. (1992). A method for disambiguating word sense in a large corpus, Computers and Humanities, 26, p. 415-439.