# SEMI-SUPERVISED TOPIC CLASSIFICATION FOR LOW RESOURCE LANGUAGES

*Daben Liu, Sam McVeety, Rohit Prasad, Prem Natarajan*

BBN Technologies, 10 Moulton Street, Cambridge, MA 02138

**ABSTRACT**

In this paper, we present a novel methodology for rapidly developing a topic-based document classification system for a language that has limited resources. Our approach, a hybrid one, combines supervised and unsupervised topic classification techniques. Given that access to native speakers is fairly limited for low resource languages, our approach requires annotating only a few broad "root" topics in the corpus. Next, unsupervised topic discovery (UTD) technique is used to automatically determine finer topics within the root topics. Lastly, we use the recently developed unsupervised topic clustering technique to organize the corpus into a hierarchical structure that enables browsing documents at multiple levels of granularity. Recognizing the need for reducing false alarms during runtime, we describe rejection techniques for discarding off-topic documents.

***Index Terms***— unsupervised topic discovery, topic clustering, Hidden Markov Model, off-topic rejection, Malay

## 1. INTRODUCTION

Topic classification, as a research area, has been widely studied for commonly used languages, such as English. Such classification has been applied to multiple domains including formal broadcast news [1] and informal Newsgroup messages [3]. However, less common languages, such as Malay [6] have not received the same attention. Therefore, there are limited resources to build a topic classification system in Malay. On the other hand, the number of Web documents in Malay language increases exponentially in the similar way as the expansion of the overall World Wide Web in general. It is estimated that there are 215M Malay internet users [5], whose need for an efficient topic classification system is no less than that for an English user.

Training a topic classification system usually requires a large corpus of documents, manually annotated with thousands of distinct topics. Such human annotation of topics is not only costly, but also likely to be incomplete due to the large set of possible topics. In addition, due to evolution of stories and occurrence of new events, a subset of pre-defined topics becomes obsolete over time, which further renders human annotation to be inadequate.

Unsupervised Topic Discovery (UTD) was proposed to automatically determine topics from a document corpus [2]. The UTD approach was extended further to automatically cluster documents based on the discovered topics [3]. This new technique, referred to as unsupervised topic clustering (UTC), organizes topics and associated documents in a hierarchical structure, and therefore enables browsing a large corpus at different levels of granularity. To prevent the topic tree from becoming unwieldy, we imposed several constraints in the topic discovery and topic clustering process. In this paper, we postulate that some amount of human supervision in constructing the hierarchical topic tree will result in a more effective organization of topics and documents. We introduce a novel hybrid approach to topic classification, which combines supervised classification and unsupervised topic classification techniques. We demonstrate the efficacy of this approach in rapidly configuring a topic-based categorization for documents in Malay language (*Bahasa Melayu*).

## 2. TOPIC BASED DOCUMENT CATEGORIZATION

Our approach leverages several topic-based document categorization techniques including hidden Markov model (HMM) based topic classification with OnTopic, UTD, and UTC. In this section we briefly review these capabilities.

### 2.1. HMM-based topic classification using OnTopic

Our topic classification engine, OnTopic[1] uses an HMM to model multiple topics in documents explicitly. The model topology is shown in Figure 1. Each topic is represented by a 1-state HMM. In addition, there is an HMM for the General Language. The probability of a word given a topic, $P(W_n|T_j)$, is associated with each topic state. *Set* is the hypothesized list of topics for the input test document.

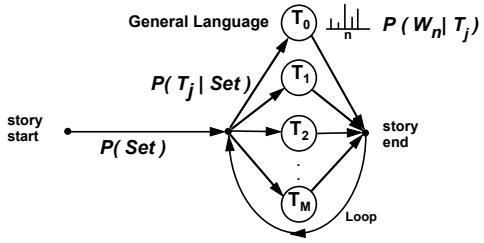


**Figure 1: Generative Model used in HMM Topic Classifier.**

The parameters of the model shown in Figure 1 are estimated using the expectation maximization (EM) algorithm that maximizes the posterior probability $P(Set|D)$, where *Set* is the list of topics labeled for document *D*.

Classification of a test document *D* is performed by considering each topic independently using the equation below, to choose a small set of likely topics:

$$\log p(T_j \mid D) = \log P(T_j) + \sum_t \phi\left\{\log\left[P(T_j \mid j \in set)^{\beta} \frac{P(W_t \mid T_j)}{P(W_t)}\right]\right\}$$

where $P(T_j|D)$ is the posterior probability of topic $T_j$ given the document D, $P(T_j)$ is the *apriori* probability for $T_j$, $P(T_j|j \in Set)$ is the average percentage of words generated by topic state $T_j$ given it is in the set of topics, and $\beta$ is an exponential weight to counteract for the independence assumption. $\phi(x)$ is equal to x when x is positive and 0 when x is negative. Topics that result in the top-N $P(T_j|D)$ values are chosen as the classified topics. In all our experiments reported in this paper, N = 3.

### 2.2. Off-Topic Rejection

In many topic spotting applications, especially where human review of in-topic documents is required, it is essential to ensure low false alarms. In [4], we presented a novel rejection mechanism based on the assumption that the General Language state in the OnTopic model can serve as the alternate model for all topics that are not of interest. This approach was inspired by the application of universal background model (UBM) in open-set speaker verification problems [9]. The algorithm can be described as follows. For each document $D$, a relevance score $s_{T_j}(D)$ is computed, and the rejection decision $\theta$ is determined by:

$$\theta(D) = \begin{cases} \text{Accept } D & \text{if } s_{T_j}(D) > \alpha \\ \text{Reject } D & \text{Otherwise} \end{cases}$$

where $T_j$ is the top-choice topic of $d$, $\alpha$ is a topic-independent rejection threshold. $s_{T_j}(D)$ is the relevance term computed as:

$$s_{T_j}(D) = \log(\frac{P(T_j \mid D)}{P(GL \mid D)})$$

where the posterior of General Language, *P(GL|D),* is used as a normalization term.

**2.3. Unsupervised Topic Discovery**

UTD attempts to automatically assign topics to a large collection of documents in order to avoid costly human annotation. As proposed in [2], no training other than the input corpus is required in the process. Each document is assigned one or more topic categories and each topic category may be assigned to multiple documents, i.e., a many-to-many mapping between the documents and topics. Here we provide a brief description of the UTD algorithm:

1. Select a few (typically five) key terms (words or phrases) for each document, where key terms are the top five terms ranked by their *TFIDF* score.
2. Prune the key term *w* if it is not a key term for more than Q documents. The surviving key terms over all document collection form the initial topic categories.
3. Train topic models with HMM-based topic model training as described in section 2.1.
4. Re-assign topic labels to the input documents by running the HMM-based topic classification.

Step 4 ensures that topic label can be assigned to documents where the label terms may not appear in the documents.

**2.4. Unsupervised Topic Clustering**

UTC extends the UTD capability by exploring the underlying relationship among topics of different granularity [3]. Using a clustering technique, UTC constructs a hierarchical tree to represent the UTD topics. Specifically, we use the following agglomerative clustering procedure for clustering topics:

1. Each UTD topic is initially assigned to its own cluster.
2. For every pair of clusters, we compute the distance between the two clusters in the pair using an appropriate distance metric. The distance measure we used is the *Mutual Information*(MI) of two topics $T_1$ and $T_2$:

$$MI(T_1, T_2) = P(T_1, T_2) \log[\frac{P(T_1, T_2)}{P(T_1)P(T_2)}]$$

   where co-occurrence probability $P(T_1,T_2)$ is computed as a ratio of the number of documents with $T_1$ and $T_2$, over the number of documents with $T_1$ or $T_2$. $P(T)$ is the probability of topic $T$.
3. The two clusters that are closest (highest MI) to each other are merged into a single cluster.
4. Steps 2 and 3 are repeated iteratively until the distance between the closest pair is higher than a threshold.

In practice, a deep tree requires more user clicks to navigate from root to leaves. To make the navigate more efficient, we introduced a tree depth penalty term to avoid a deep and unbalanced tree:

$MI(T_1, T_2) / Depth^{\gamma}$ where $\gamma$ is a configurable factor.

## 3. SEMI-SUPERVISED TOPIC CLASSIFICATION

Our proposed approach combines both supervised topic classification and unsupervised topic discovery techniques for topic-based categorization in a new, low resource language. Using limited and affordable human annotation, supervised techniques provide the first partition of topic space so that unsupervised techniques can work within a confined range and produce more manageable results.

Before going into details of the system design, we first define the notion of "*root topic*", which is essential in introducing human supervision before unsupervised topic discovery and clustering.

**3.1. Root Topics**

Root topics are topics designed to cover a broad category of documents and be applicable across languages and cultures. Specifically, we identify three distinctive properties that govern whether a topic is a root topic:

1. *Broadness/General*. A root topic must be broad and cover a wide range of documents with more detailed topics. "Economy" would be a root topic as opposed to "Alan Greenspan."
2. *Persistency*. A root topic should be relatively insensitive to time. Topics such as person name may become obsolete once the person is no longer active in news stories. However, "Education" or "Economy" topics are persistent topics.
3. *Applicability*. A root topic should be applicable for most languages or cultures. This is one way to validate the "Broadness". Consequently, the *same* list of root topics can be used to bootstrap development in other languages.

The goal for defining root topics is twofold. First, root topics enable efficient topic annotation since the list of pre-defined topics is relatively short. Second, unsupervised techniques can be applied within each root topic so that the resulting topic tree is more manageable in size and supports efficient navigation.

In our experiments, we used the English PSM documents [1] to bootstrap the list of root topics for Malay documents. The PSM data contains 95K broadcast news articles with 7002 manually labeled topics. Under the assumption that root topics tend to be general and frequent, we sorted the topics in the PSM corpus based on document frequency. Then, we selected the root topics from the top of the frequency sorted list using the following three criteria:

1. Discard any geography-specific topics, such as "California".
2. Discard the topic if it is covered by another broader topic of higher frequency. For example, since "Terrorism" is already a root topic, "bombing" was discarded
3. Merge topic labels that are most likely to co-occur in a document. For example, "Internet" and "Computer" are merged into one root topic "Internet, Computer"

We derived an initial list of 37 root topics using the above procedure. This list was further refined during topic annotation process of the Malay corpus by a native speaker of Malay (more in Section 5). The Malay annotator was instructed to include new root topics if he found many documents that cannot be categorized by any of the root topics in the initial list. The annotator added 4 new root topics, while annotating the first 500 documents. The 4 topics were: Royalty, Agriculture, Event/Announcements, and Travel/Tourism. It is worth noting that no additional topics were added after annotation of the first 500 documents. This suggests that the resulting 41 root topics cover broadly the documents in the corpora. As examples, "Education", "Economy", "Law, Court" are among the selected root topics.

**3.2. System Design and Architecture**

Figure 2 shows the system architecture design of our proposed approach. The topic classification system is developed by the following steps:

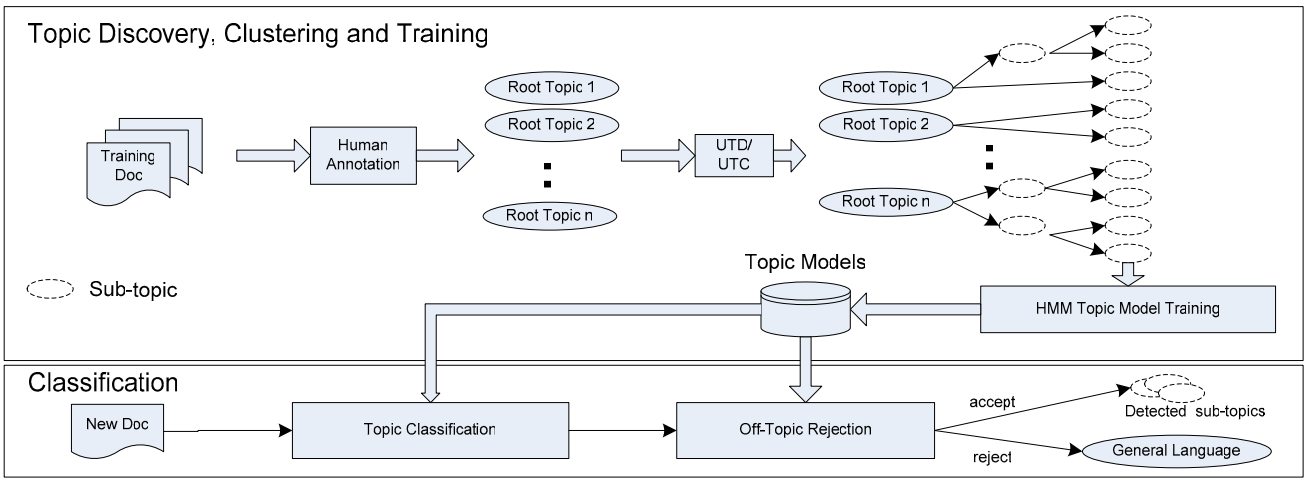


**Figure 2. System Diagram for Semi-Supervised Topic Classification System.**

1. Collect training data from public and open source domain.
2. Manually label root topics for each training document.
3. For each root topic, a detailed topic list is automatically discovered by applying UTD to all the documents associated with that root topic.
4. UTC is applied to the discovered topic list to create a hierarchical topic tree. Each leaf node is one UTD topic. The topic tree provides an efficient hierarchical organization of otherwise unstructured documents.
5. HMM topic models are trained with the leaf-node topics, and are used for topic classification and off topic rejection.

The result of topic classification of a test document is either one or more topics in the topic tree, or a General Language label indicating an off-topic document.

As one can see from the description of our procedure, even though we use Malay language for our experiments to demonstrate our methodology, the same technique can be applied to any other languages that have limited resources.

## 4. MALAY LANGUAGE AND RESOURCES

Historically, Malay language has been written using various types of scripts. For our experiments, we used Malay documents written in Latin alphabet called Rumi. Rumi is the widely adopted script for both formal and colloquial writings. Since the Rumi script uses the same set of characters as English, no special encoding is required to process electronic documents in Malay language.

For our data collection, we first searched the Internet to locate Malay resources that can be useful for topic classification, such as text documents, annotated corpora, Malay stemmers, topic lists, stop word lists, or lexicons, etc. Although a large number of unstructured documents are available on the Internet from major news sites, no topic-annotated data was located. We did find some research papers about Malay-specific stemming [7]. However no stemmer tool is publicly available as the Porter's Stemmer for English. A stop list from New York University's GMA Language Resources [8] was downloaded for improving the topic classification process.

## 5. DATA COLLECTION AND ANNOTATION

We identified three websites as sources for mainstream Malay news articles: {www.bernama.com.my, www.utusan.com.my, www.bharian.com.my}.All three have archived documents dated back to 2002. We developed our own web harvesting tool that uses user-defined regular expression to select content pages, while discarding advertising or table-of-content pages. Perl HTML::Parser 3.0 was used to extract the textual contents of the downloaded HTML pages. Using the tool, we were able to download 46K documents with diverse content within 3 days.

To annotate root topics, we developed an annotation tool with graphical user interface (GUI). A native Malaysian was hired and trained to do the annotation. The annotator was able to memorize all the 41 root topics within the first few hours, and found it efficient to select the topics for the documents using the annotation tool. The annotation speed on the average was 75 documents per hour. We annotated 5274 annotated Malay documents, with an average 2 root topics for each document.

## 6. EXPERIMENTAL RESULTS

### 6.1. Supervised Topic Classification

The root topic annotation provides a ground truth for test documents. Therefore, in this section we first measure the accuracy of supervised topic classification with OnTopic trained on Malay documents. The accuracy metric we used is the top-1 topic accuracy, which is defined as the percentage of times the top-choice topic was the correct answer.

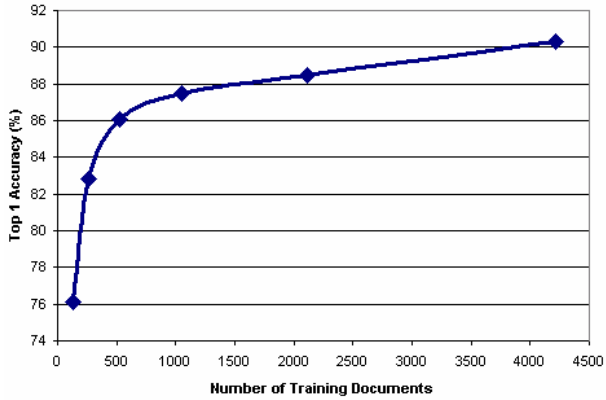


**Figure 3. Top-1 Accuracy on Malay Documents.**

We randomly divided the annotated set of 5274 documents into 80% training data and 20% test data. To investigate the effect of number of training documents on the accuracy, we conducted a series of experiments with each experiment doubling the amount of training documents from the previous experiment. Figure 3 shows

the experiment results. As we can see from Figure 3, the accuracy improves significantly up to 500 training documents, and then starts to taper off. However, even at the full amount of training data, there is still potential upside trend for further improvement. With all training documents, the top-1 accuracy is 90.5%. For comparison, the top-1 accuracy for the English AFE newsgroup messages is 91.2% [3].

### 6.2. Off-Topic Rejection

Experiments in the previous section used all "closed-set" data, that is, all documents classified were of topics of interest. To test rejection performance, we simulated the off-topic scenario by removing some of the root topics and creating an "open-set" corpus. As shown in Table 1, this results in 3 sets of documents.

| SET | ON-TOPIC | OFF-TOPIC | # DOC | # TRAINING | # TEST |
|---|---|---|---|---|---|
| ON | YES | NO | 3643 | 2443 | 1200 |
| MIX | YES | YES | 968 | 668 | 300 |
| OFF | NO | YES | 663 | 463 | 200 |

**Table 1. Training/Test partition across categories.**

The set labeled "ON" consists of all documents that have only topics of interest; "OFF" consists of documents that are not on topics of interest; "MIX" documents contain documents that have both of topics of interest and some topics not of interest. In practice, all three types of documents could occur in test conditions. For test purposes, we consider ON and MIX documents as *on-topic*, while OFF documents are considered off-topic and should be discarded.

We performed rejection experiments using the procedure described in Section 2 and detailed in [4]. Three sets of OnTopic models were trained using ON, ON+MIX, and ON+MIX+OFF, respectively. The receiver operating characteristic (ROC) curve is shown in Figure 4. As shown in the figure, adding off-topic documents, which is used for General Language model training, significantly improves the performance. Adding "MIX" data, which has both on topic and off topic contents, the True Positive score degrades, especially at low false alarm rate. The equal-error-rate for the best ON+MIX experiment is 12.5%.

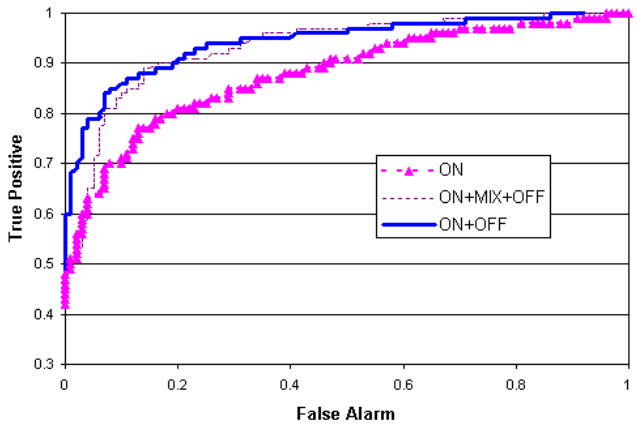


**Figure 4. ROC Curve for Off-Topic Document Rejection.**

### 6.3. Unsupervised Topic Clustering for Malay

UTD was performed on documents from each of the 41 root topics. During the first round, UTD produced 2699 automatically discovered topics. Next, the Malay annotator identified topic labels that were not really topics. 52 such terms were found, e.g. ITU (that), SEBANYAK (as many as), etc. We added all 52 terms to the stop list, and performed UTD again. This time 2637 UTD topics were discovered, and all were deemed as valid topics.

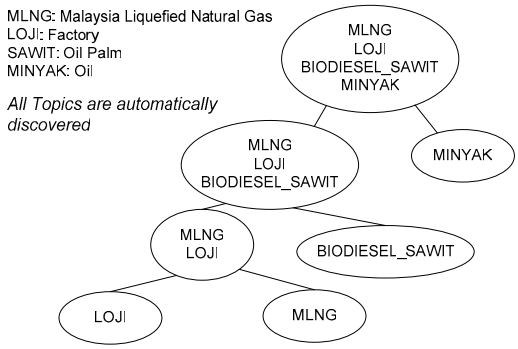


**Figure 5. Example of a Sub-Tree for Root Topic "Economy".**

Next, hierarchical topic trees were created for each root topic using UTC [3]. The tree depth ranged from 2 to 8, which means at most 8 clicks are required to traverse one full tree. Figure 5 shows a typical sub-tree from root topic "Economy." Documents from this tree are on alternative bio-fuel sources to replace fossil oil.

## 7. CONCLUSION

In this paper, we demonstrated a semi-supervised methodology for topic-based categorization of Malay documents. The same methodology can be applied to any other language, including languages with limited resources. The classification system developed for Malay was trained on open-source data harvested from the Internet. We also showed that root topics are extremely effective in improving the efficiency and accuracy of human annotation. The supervised topic classification results on Malay are at par with that of state-of-the-art systems for categorizing English documents. Unsupervised topic discovery and clustering were used to automatically create a hierarchical organization of topics and associated documents. Subjective evaluation of the topic tree with a Malay speaker suggests that the topic tree is effective in browsing a new corpus of Malay documents.

## 8. REFERENCES


[1] Schwartz et al., "A Maximum Likelihood Model for Topic Classification of Broadcast News," *Proc. Eurospeech*, Greece, 1997.

[2] S. Sista et al., "An Algorithm for Unsupervised Topic Discovery from Broadcast News Stories," *Proc. ACM HLT*, San Diego, 2002.

[3] R. Prasad et al., "Finding Structure in Noisy Text: Topic Classification and Unsupervised Clustering," *Proc. AND*, 2007.

[4] K. Subramanian et al., "Optimal Estimation of Rejection Thresholds for Topic Spotting," *ICASSP'07*, Hawaii, USA, 2007.

[5] M. Hamzah et al., "On retrieval performance of Malay textual documents", *Proc. of the 24th IASTED AIA'06*, February 2006.

[6] Z. Yusoff et al., "Computational linguistics in Malaysia," *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hong Kong, Pages: 1 – 2, 2000.

[7] S. Y. Tai et al., "On designing an automated Malaysian stemmer for the Malay language," *Proceedings of the fifth international workshop on Information retrieval with Asian languages,* Pages: 207 – 208, Hong Kong, China, 2000.

[8] A. Argyle et al. "http://nlp.cs.nyu.edu/GMA/docs/resources.html," *GMA Language Resources*, 2004.

[9] D. Reynolds et al., "Speaker Verification using Adaptive Gaussian Mixture Models," *Digital Signal Processing*, pp.19-41, 2000.