Discovering Thematic Knowledge from Code-Mixed Chat Messages Using Topic Model

Kavita Asnani[1], **Jyoti D. Pawar**[2]
[1]Goa College of Engineering, Goa, India
[2]Goa University, Goa, India
E-mail: kavita@gec.ac.in, jyotipawar@gmail.com

Abstract

In current times, the trend of mixing two or more languages together (code-mixing) in communication on social media is very popular. Such code-mixed chat data is enormously generated and is usually noisy, sparse and exhibits high dispersion of useful topics which people discuss. In such a scenario, it is very challenging to automatically extract relevant thematic information which contributes to useful knowledge. In order to discover latent themes from multilingual data, a standard topic model called Probabilistic Latent Semantic Analysis (PLSA) is used in existing literature. However, it addresses the inter-sentence multilingualism. In this paper, we propose a novel method which is basically based on co-occurrences of words within a code-mixed message. Thus built co-occurrence matrix for chat is exposed to PLSA which is used to discover thematic knowledge from it. In such code-mixed chat text, inter-sentence, intra-sentence and intra-word level code mixing may randomly occur. We have proved with extensive experiments that it is possible to use this strategy to discover latent themes from semantic topic clusters. We tested our system using FIRE 2014 dataset.

Keywords: Thematic Knowledge, Code-mixed data, Topic model, PLSA

1. Introduction

In recent years, communication over social networking has become very popular. Therefore, most of the research on social media text has concentrated on English chat data or on multilingual data at inter-sentence level where each message is monolingual. However, majority of chat communication now occurs in random mix of languages (Jamatia et al., 2015). (Chandra, 2014) presented a study to identify the language mixing pattern in Bolywood movies songs. From 3784 Hindi songs of 1008 movies he found that 1,38,146 unique words were extracted out of which 2383 were unique English words. His analysis claimed that the mixing of English in songs is popularly increasing with time. Code mixing occurs when a person changes language (alternates or switches code) below clause level, internally inside a sentence or an utterance (Jamatia et al., 2015). In particular, India is a multilingual country having great influence of code-mixing in communication. (Das and Gamback, 2014) reported code-switching in Facebook chat messages mixed in English-Bengali or English-Hindi, and stated that inter-sentential switching account for 60.23% and 54.71% respectively. Also, intra-sentential switching account for 32.20% and 37.33% respectively. Thus, code-mixing while chatting has become prevalent in the current times. However, such large volumes of short and long chat messages contain lot of noise and have the main themes of discussion dispersed. We believe that the thematic knowledge from such data could point to relevant topics of interest to the chat system administrator or user. Unfortunately, it is not an easy task as the messages often could be code-mixed in multiple languages at different levels of code complexity. In this work we try to address these challenges based on the hypothesis which states that the words co-occurring in the similar context tend to be semantically similar.

(Chandra and Kundu, 2013) proposed a hybrid approach combining rule based and statistical based method for language identification in code-mixed text. By automatic detection of English words in Benglish and Hinglish text, he pointed out challenges in computational analysis of code-mixed sentences like difficulty in machine translation, Cross-Lingual Information Retrieval (CLIR), POS tagging and ambiguities in mixed words. Due to the difficulties and lack of available language identification systems, we propose to drop the structure of messages by breaking them into bag of words and representing them in a co-occurrence matrix, thereby skipping the need of language identification. We present a novel approach based on Probabilistic Latent Semantic Analysis (PLSA) which is capable of extracting latent thematic knowledge from code-mixed chat messages.

The remainder of this paper is organized as follows: section 2 presents related work, section 3 describes the proposed model for thematic knowledge discovery using PLSA, section 4 gives experimental evaluation and section 5 states conclusion and future work.

2. Related Work

Topic models are powerful tools to identify latent text patterns in standard text domains like web page citation network; but social media text differs completely (Hong and Davison, 2010). Content analysis in social media like Twitter, poses unique challenges as posts are short and in any language unlike the standard written English on which many supervised models in machine learning and NLP are trained and evaluated (Ramage et al., 2010). Topic mixture for both messages and authors in the twitter corpus was inferred by (Hong and Davison, 2010). They used topic modeling for predicting popular twitter messages and classifying twitter users and corresponding messages into topical categories.

(Huang et al., 2013) proposes multi-task multi-label (MTML) classification model that combines sentiment and topic classification of tweets. They stated that as tweets are short, noisy and written in informal language they make classic methods of natural language processing not well applicable. Also, topics of tweets may not be perfectly exclusive and content of a tweet may cover multiple topics. They mapped each tweet separately as a feature vector. They applied Maximum Entropy (ME) to obtain probabilistic classification of both sentiments and topics concurrently.

(Mcauliffe and Blei, 2008) proposed supervised topic models which functions primarily on prior knowledge and assumes the prior knowledge to be correct. **(**Ramage et al., 2010) proposed Labeled LDA which employs supervision on LDA that performs content analysis and classification of twitter feeds to characterize users by the topics they most commonly use. We cannot use supervised techniques as they need to prior classify messages into predefined classes. This requires good prior knowledge about the data which is not feasible in our case as code-mixed chat data is generated randomly in any language. An unsupervised topic model is our preference as they do not need prior knowledge about data to infer latent themes from the text collection. Topic models such as PLSA (Hofmann, 1999) have been successfully applied to many applications such as sentiment analysis as they do not use any prior knowledge or external resources (Titov and McDonald, 2008).

(Balahur and Turchi, 2013), presented a method to perform sentiment analysis on multilingual tweets. They claimed that it is challenging to process tweet data as it is multilingual and contains slang, emoticons, repetition, misspellings etc. To address this, they built a system processing tweets in English taking into account specificity of expression and then using a standard machine translation system translated the data from English to four languages- Italy, Spanish, French and German. Their work essentially needed a language identifier that separated the data from different languages. Further they manually corrected the test data and created gold standard for each of the target languages.

In a multilingual country like India, we have around 22 official languages across 29 states and millions of people communicating over social networks for routine tasks. Thus, our work is motivated by the ever increasing occurrence of complex code-mixing resulting in large volumes of chat text having useful knowledge highly dispersed in large noise. Hence, our proposed approach, attempts to verify the claim that probabilistic topics can be used for thematic knowledge discovery for the chat user or administrator.

3. Thematic Knowledge Discovery using PLSA

A topic model takes as input a set of documents, and generates clusters of words called 'topics'. These topics help to extract themes underlying a dataset. A popular topic model by (Hofmann, 1999) called Probabilistic Latent Semantic Analysis (PLSA) is an unsupervised model. This model takes as input the value 'k' as the number of topics. PLSA takes as an input a dataset and models two kinds of distributions: i) a document-topic distribution that determines the distribution of topics within a document, and ii) a word-topic distribution that determines the distribution of words across the topics. The two distributions are estimated using an Expectation-Maximization approach. The output of PLSA is the estimation of top 'n' relevant words for each topic.

Understanding of social media text especially when mixing of multiple languages occurs at sentence level or even word level in dynamically growing noisy messages is a very complex task. In our proposed approach, we model co-occurrences of words in a message, as a unit and then we use probabilistic topic model PLSA (Hofmann, 1999) to obtain useful thematic representation of our data. Our proposed method is designed taking into account code-mixed English-Hindi chat data. The plate notation for our proposed PLSA based model is shown in Figure 1. We have presented our complete method in Algorithm 1.

Each code-mix chat message m and collection of messages M in the figure 1 is represented as an entity as expressed in equation 1.

$$M = \{m_1, m_2, m_3, \ldots, m_n\} \text{ where } m \in M \quad \text{----------(1)}$$

We represent collection of such message entities as bag-of-words over the wide chat vocabulary V shown in figure 1, and expressed in equation 2. Topic models commonly represent data as bag-of –words (BOW), which means that the ordering of words is not considered. This characteristic is suitable in our context as we are dealing with code-mixed data, so by BOW technique, structure is dropped and hence each word is treated independently. As a result, there is no need to consider the language in which the word is written. Therefore, our proposed method handles random code-mixing as we do not perform language identification at all. The emphasis is only to find if the word is belonging to a certain topic with high probability. Eventually, the words which do not contribute to a topic will be treated as insignificant words which do not essentially represent useful information.

$$m = x_1, x_2, x_3, \ldots\ldots, x_{|m|} \quad \text{----------(2)}$$

where $x_i \in V = \{ w_1, w_2, \ldots, w_m\}$ is a word.

The key step in our method is to determine context and for that we believe in "higher-order co-occurrence", i.e., how often words co-occur in same contexts (Heinrich, 2009). The word by message matrix is constructed by computing the frequency of each word in the respective message

using updateCoOccurenceMatrix(m_i, w_i) in the Algorithm 1. As the vocabulary of code-mixed chat data across languages generate hundreds or thousands of distinct words, the co-occurrence matrix becomes large. We eliminate the least significant noise words by using a stop word list.

In order to extract latent thematic information we employ PLSA topic language model which takes as a parameter number of topics k giving Z set as in equation 3.

$Z = \{z_1, z_2, z_3, \ldots z_k\}$ where $z \in Z$ is a topic. ----------(3)

Now, following the probabilistic topic based language model, we assume the following:

i) Every code-mixed message m_n is selected with probability P(m)

ii) Every topic z_i is chosen from a mixture of latent topics in that message m_n with probability $P(z_k \mid m_n)$ and $z_1 + z_2 + z_3 + \ldots + z_n = 1$

iii) Every word w_i in the message m_n is chosen from multinomial topic distribution with probability $P(w \mid z_i)$.

Since every topic is distribution over words and every message is distribution over topics, words and messages are conditionally independent. The same is specified giving joint probability in the equation 4.

$$P(w,d) = \sum_{z \in Z} P(z)P(m|z)P(w|z) \quad \text{----------(4)}$$

Therefore, objective function of PLSI is expressed in the equation 5 as,

$$L = \prod_{m \in M} \prod_{w \in W} P(w|m)^{n(m,w)} \quad \text{----------(5)}$$

Since this gives non-convex optimization problem log is done as shown in equation 6.

$$\ell = \log L \quad \text{----------(6)}$$

$$= \sum_{m \in M} \sum_{w \in W} n(m,w) \log \sum_{z \in Z} P(w|z) \cdot P(z|m)$$

Since we want to select a distribution that gives a word higher probability P(w|z), Expectation Maximization (EM) algorithm is used by performing the following steps:

1. Initialize P(w|z), P(m|z) and P(z) with random values using rnd_init() function in Algorithm 1.
2. Iteratively update them using E-step and M-step given in equation 7 to 10.
3. Stop when the likelihood ℓ given in equation 6 does not change.

In E-step we guess the latent values z. It does the job of augmenting the messages and words with z information as expressed in equation 7.

$$P^{(n)}(z|w,m) = \frac{P(z)\, P^{(n)}(m|z)\, P(w|z)}{\sum_{z'} P(z')\, P(m|z')} \quad \text{----------(7)}$$

M-Step step takes advantage of inferred z values and groups words that are in the same distribution as expressed in equation 8, 9 and 10.

$$P^{(n+1)}(w|z) = \frac{\sum_m n(m,w)\, P^{(n)}(z|m,w)}{\sum_{m,w'} n(m,w')\, P^{(n)}(z|m,w')} \quad \text{----------(8)}$$

$$P^{(n+1)}(m|z) = \frac{\sum_w n(m,w)\, P^{(n)}(z|m,w)}{\sum_{m',w} n(m',w)\, P^{(n)}(z|m',w)} \quad \text{----------(9)}$$

$$P(z) = \frac{\sum_{m,w} n(m,w)\, P^{(n)}(z|m,w)}{Q} \quad \text{----------(10)}$$

where $Q = \sum_{m,w} n(m,w)$

EM iteratively improves our initial estimate of parameters by using E-step and then M-step. E-step is to compute the lower bound (latent variable value) and M-step is to maximize the lower bound. Since our data is dynamically growing and very noisy our immediate objective is to extract relevant themes which could express meaningful context.

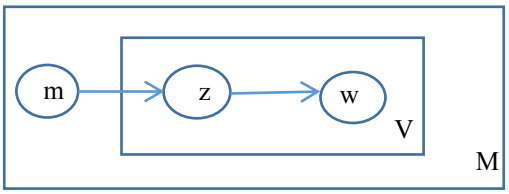

Figure 1: PLSA Plate Notation for Code-mix Messaging System

Algorithm 1 Constructing Topic-based Aspect Clusters

Input : Code-mixed chat message collection M, k, n
Output : Top k Thematic clusters

1. for each message $m_i \in$ M do
2. for each word position $w_i \in m_i$
3. $M_V \leftarrow$updCoOccurenceMatrix(m_i, w_i);
4. endfor
5. endfor
6. for each topic $z_i \in$ Z do
7. for each `message $m_i \in$ M do
8. for each word position $w_i \in m_i$ do
9. $P(z_i | w_i, m_i) = 0$;
10. $P(w_i | z_i)$, $P(m_i | z_i)$, $P(z_i) \leftarrow$ rnd_init();
11. endfor
12. endfor
13. endfor
14. repeat
15. update P(z|w,m); *//Apply E-step using M_V*
16. update P(w|z), P(m|z), P(z); *//Apply M-step using M_V*
17. update ℓ;
18. until nochange(ℓ);

```
19. for each topic z_i ∈ Z do
20.   for each message m_i ∈ M do
21.         for each word position w_i ∈ m_i do
22.               score_{w_i} ← P(w_i | z_i);
23.               w_i ← w_i + score_{w_i}
                        //Augment each w_i with its score
24.         endfor
25.     endfor
26.  endfor
27. repeat
28.   for each topic z_i ∈ Z do
29.         T_c ← sort( w_i);
30.    endfor
31. until (k , n)          //Sort k clusters with top n words
32. return T_c
```

4. Experimental Evaluation

4.1 Dataset

For discovering latent themes, we performed experiments on FIRE 2014(Forum for IR Evaluation)[1] shared task on transliterated search; which comprises of data from English mixed with six other Indian languages. The English-Hindi corpora from FIRE 2014 was introduced by (Das and Gamback, 2014), and it consists of 700 messages with the total of 23,967 words which were taken from a Facebook chat group for Indian University students. As compared to the other language pairs in the corpora, the said English-Hindi corpus had as high as 80% of code-mixing percentage due to the frequent short-hand language or slang used in the two languages randomly during the chat(Das and Gamback, 2014). For such code-mixed text it is highly desirable to have a means of automatic discovery of latent thematic knowledge.

4.2 Pre-processing

We performed tokenization of the input message text and then removed the stop words[2] and punctuations. We plan to consider the slang occurring in the chat text in our future work as we found it difficult to find a suitable normalization method for fixing informal abbreviations in chat data in Hindi language. We observed from the messages in our experimental corpus that the slang within the messages is likely to recur consistently than across the messages e.g. the word "great" used as "gr8" consistently in the same message. Since, our proposed method considers a message as a document; and words which co-occur with similar probabilities belong to the same topic and rejects words that have different probabilities across topics; we would not expect slang to bias the results as such but will affect the coherence of topics.

4.3 Code-Mixed Message as a Document

As stated in (Titov and McDonald, 2008), topic models are applied to documents to produce topics from them. Since our aim is to discover themes from chat messages, we treat each message independently and divide it into stream of words. Although, relationship between messages is lost, the data in BOW across the vocabulary of chat messages contributes to the construction of the co-occurrence matrix. This representation is fair enough as it eliminates the need for language identification across languages code-mixed in a message.

4.4 Effects of Thematic Knowledge

In order to analyse the performance of our method with respect to topic numbers k, we experiment with different values and observe the effect of the same. Each topic was displayed as a list of words, sorted in the decreasing order of probability of that word belonging to the topic. We tested the performance of the proposed system by evaluating the interpretability of topics and analysed if they conform to human knowledge. We worked with two judges who had experience in chatting on social networking sites. Thematic clusters obtained as output by the proposed method are rankings based on word probability, thus in order to know the number of correct topical words, we evaluated these rankings using Precision at different levels n, where n is the rank position, as used in (Zhao et al., 2010). We performed this evaluation in two steps:

i) Topic Annotation and Evaluation

We followed (Mimno et al., 2011) (Chuang et al., 2013) to evaluate quality of each topic as ("good", "intermediate", or "bad"). The topics were annotated as "good" if they contained more than half of its words that could be grouped together thematically, otherwise "bad". Each topic was presented as a list of 20 most probable words in descending order of probabilities under that topic. The human judges were unaware of the model which generated the topics. For each topic, the judges annotated the topics independently and then we aggregated their results. Table 1 reports the Cohen's Kappa score for topic annotation, which is above 0.5, indicating good agreement. We observed a high score at k=3 due to consistently contributing few topics resulting in strong agreement. According to the scale the Kappa score increases with more number of topics as the topics get thematically stronger.

Table 1: Cohen's Kappa for inter-rater agreement

Index	1	2	3	4	5	6	7
k	3	6	9	12	15	18	21
Precision @k	0.92	0.554	0.627	0.639	0.665	0.722	0.736

[1] http://www.isical.ac.in/~fire/

[2] https://sites.google.com/site/kevinbouge/stopwords-lists

ii) Topic Size and Evaluation

We followed the instructions in (Mimno et al., 2011), and as a baseline we consider the effect of the topic size for evaluating the topic quality. Again we consider each topic to be a cluster of top 20 probability words of the output of our proposed method. Topic size refers to the number of tokens assigned to each topic. We requested human experts to provide annotations in the same scale as ("good", "bad", "intermediate") for each word in the topic manually. Since judges had already annotated the topics earlier, annotating words in a topic was not difficult. We evaluated the topic quality by computing the coherence score and we assign it rating as suggested in (Chuang et al., 2013) as {1, 0.5, 0} for each ("good", "bad", "intermediate") response respectively. We calculate the coherence score using the equation 11.

Coherence score = ((#of good topics*1) + (# of intermediate topics *0.5))/total # of words ----------(11)

Table 2 shows few example topics derived with the respective coherence score.

Table 2: Example Topics with different coherence score (High coherence indicates better thematic knowledge)

0.725	gandhi years indian din citizenship india father Italian education make toilets family studied born minister power officially indira cries ek
0.665	hai dont people understand police traffic mentality things sold rapes called toh laws realize Mumbai india dear made
0.525	toh love na ki india coz ish hui karna kya english kro agree letter yr politicians rahul gandhi khud don't agar

We can see from Table 3 that the coherence score increases with the increase in the topic size. As the number of the words per topic increase the coherence score also increases. Our topic coherence score is indicative of our observation that for the topics having high coherence score have more than half of the words that are annotated "good" and such are the words which commonly co-occur in co-occurrence matrix. For instance in "good" topics most of the words are either "good" or "intermediate" and such words are highly co-occuring.

Table 3 Association between topic size and coherence

Topic	Size	Coherence Score	Coherence Score
3	5	0.3	0.266667
	10	0.358333	0.325
	15	0.388889	0.363889
	20	0.454167	0.4125
6	5	0.288889	0.466667
	10	0.338889	0.533333
	15	0.388889	0.585185
	20	0.411111	0.583333

Based on the evaluation results we want to highlight the following points:

1. Coherence score is high when numbers of "good" or "intermediate" words are high. Therefore, good words contribute to thematic topical words.
2. "Bad" words are due to noise and repeated ungrammatical or slang words which co-occur.

5. Conclusion

This paper proposed a novel task of discovering thematic knowledge from code-mixed chat text by computing co-occurrences between the words across the languages and utilizing the PLSA topic model for extracting latent topics. We conducted experiments on facebook chat data from FIRE 2014 and demonstrated that our method is applicable for discovering latent themes at different granularity levels. In our future work, we want to implement our method after applying an optimized normalization on the code-mixed data for obtaining more precise themes.

References

Balahur, A. and Turchi, M. (2013). Improving sentiment analysis in twitter using multilingual machine translated data. In *RANLP*, pp 49-55.

Chandra, S. and Kundu, B. (2013). Hunting Elusive English in Hinglish and Benglish Text: Unfolding Challenges and Remedies. *In Proceedings of the 10th International Conference on Natural Language Processing (ICON-2013), at Centre for Development of Advanced Computing (CDAC), Noida.* India:

Macmillan Publishers.

Chandra, S. (2014). Main ho gaya single I wanna mingle...: An evidence of English Code-Mixing in Bollywood Songs Lyrics Corpora. *In Proceedings of the First Workshop on Language Technology for Indian Social Media Text with The Eleventh International Conference on Natural Language Processing (ICON-2014),* 18-21 December, 2014, Goa University, Goa, India.

Chuang, J., Gupta, S., Manning, C., and Heer, J., (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. *In Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp612-620.

Das, A., and Gamback, B. (2014). Identifying languages at the word level in code-mixed indian social media text. *In Proceedings of the 11th International Conference on Natural Language Processing*, Goa, India, pp169-178.

Heinrich, G. (2009). A generic approach to topic models. *In Machine Learning and Knowledge Discovery in Databases*, Springer, 2009, pp 517-532.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *In Proceedings of the fifteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1999, pp289-296.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. *In Proceedings of the first workshop on social media analytics*, ACM, 2010, pp 80-88.

Jamatia, A., Gamback, B. and Das, A. (2015). Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages.*Recent Advances in Natural Language Processing (RANLP)*, pp239-48.

Mcauliffe, J. D. and Blei, D. (2008). Supervised topic models. *In Advances in neural information processing systems*, pp121-128.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, M. (2011). Optimizing semantic coherence in topic models. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, *Association for Computational Linguistics 2011,* pp 262-272.

Ramage, D., Dumais, S.T., and Liebling D. J. (2010). Characterizing microblogs withtopic models. *In ICWSM*.

Shu-Huang, Wei-Peng, Jingxuan, L. and Dongwon, L. (2013). Sentiment and topic analysis on social media: A multi-task multi-label classification approach. *In Proceedings of the 5th annual ACM web science conference*, ACM, 2013, pp172-181.

Titov, I., and McDonald, R. (2008). Modeling onlinereviews with multi-grain topic models. *In Proceedings of the 17th international conference on World Wide Web*, ACM, 2008, pp 111-120.

Wayne, Xin-Zhao., Jiang, J., Hongfei, Y., and Xiaoming-Li. (2010). Jointly modeling aspects and opinions with a maxent-lda hybrid. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2010, pp 56-65.

Zhang, H., Wang,Chang-Dong. and Lai, Jian-Huang. (2014). Topic detection in instant messages. *In Machine Learning and Applications (ICMLA)*, IEEE, 2014, pp219-224.