# Extraction of Code-mixed Aspect Topics in Semantic Representation

Kavita Asnani, Jyoti D. Pawar

Goa University, Department of Computer Science and Technology,
Goa, India

{dcst.kavita, jdp}@unigoa.ac.in

**Abstract.** With recent advancements and popularity of social networking forums, millions of people virtually connected to the World Wide Web, commonly communicate in multiple languages. This has led to the generation of large volumes of unstructured code-mixed social media text having useful aspects of information highly dispersed. Aspect based opinion mining relates opinion targets to their polarity values, in a specific context. It is known that since aspects are often implicit, detecting and retrieving them is a difficult task. Moreover, it is very challenging as the code-mixed social media text suffers from its associated linguistic complexities. As a standard, topic modeling has a potential of extracting aspects pertaining to opinion data from large text. This results not only in retrieval of implicit aspects but also in clustering them together. In this paper we propose knowledge based **l**anguage independent **c**ode-**m**ixed **s**emantic LDA (lcms-LDA) model, with an aim to improve the coherence of clusters. We find that the proposed lcms-LDA model infers topic distributions without language barrier, based on semantics associated with words. Our experimental results showed an increase in the UMass and KL divergence score indicating an improved performance in the resulting coherence and distinctiveness of aspect clusters in comparison with the state-of-the-art techniques used for aspect extraction of code-mixed data.

**Keywords.** Code-mixed aspect extraction, knowledge-based topic modeling, semantic clustering, BabelNet, language independent semantic word association.

## 1 Introduction

In sentiment analysis, aspect extraction aims to extract attributes of entities called aspects that people express in their opinions [15]. In social media context, users actively communicate in a mix of multiple languages, thereby generating large content of code-mixed data [7]. Since such data occurs as random mix of words in different languages, context is spread across languages and therefore semantic interpretation gets difficult and can be resolved only by explicit language identification systems [4, 28]. However, it is computationally intensive and practically infeasible to find or build translation tools and parallel corpora for each language pair in the code-mixed data. Therefore, we are interested in designing an automatic process for language independent retrieval of aspects from code-mixed data, without an aid from parallel corpora or translation resources. Also, the useful topics of information are highly dispersed in the high dimensional code-mixed social media content.

Therefore there are two key challenges:

1. Extraction of useful aspects from code-mixed social media data across the language barrier.
2. Making the aspect extraction fault tolerant and coherent by addressing semantics while grouping them.

In the existing literature, two models have been highly recommended - Probabilistic Latent Semantic Analysis (PLSA), [10] and Latent Dirichlet Allocation (LDA), [5]. Both the models infer latent 'topics 'from 'document 'and 'word'. In [2], the authors proposed code-mixed PLSA which is based on co-occurrence matrix for representation of code-mixed words from chat messages and is modeled on PLSA for aspect discovery. However, the work carried out by using PLSA and LDA in monolingual context, resulted in extraction of some incoherent aspects [6]. Considering the code-mixed content, [2]

Computación y Sistemas, Vol. 22, No. 1, 2018, pp. 55–63
ISSN 1405-5546
doi: 10.13053/CyS-22-1-2771

attributed the issue of inclusion of incoherent aspects (aspects placed in incorrect topic clusters), in the output to the noisy, un-normalized and out of context words. In this paper we handled this issue in two steps: first, use of shallow parser [27], to obtain normalized output at the pre-processing stage and second, associating semantics to language independent word distributions in Latent Dirichlet Allocation (LDA) algorithm [5], for retrieval of coherent latent aspects. Therefore, the need for improvement in coherence of aspect clusters has been addressed using multilingual synsets.

In our system multilingual synsets are extracted from freely available lexicographic dictionary called BabelNet [21]. The multilingual synsets for two sample words, one in English and the other in Hindi is shown in Fig. 1, and these are coherently related in lcms-LDA based on semantic similarity.

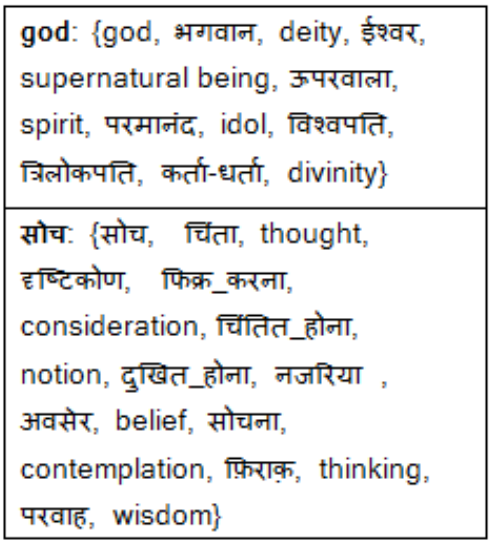


**Fig. 1.** Topic distinctivity comparison for lcms-LDA

In summary, our work makes the following contributions:

— Automatic extraction of useful aspects from code-mixed data by learning from language independent monolingual word distributions.

— Improvement in coherence of clusters by shifting from merely statistical based co-occurrence grouping to semantic similarity between words.

The structure of the paper is as follows: Section 2 illustrates related work in aspect extraction and related issues; Section 3 describes the proposed model; Section 4 presents experimental results with explanation on how aspects are extracted from code-mixed text and comparative evaluation of lcms-LDA with the state-of-the-art techniques for code-mixed aspect extraction; finally, Section 5 presents concludes the paper.

## 2 Related Work

Social media is a source of huge amount of opinion data on the web. In multilingual countries, people on social networking forums, often communicate in multiple languages, both at conversation level and at message level [14]. Majority of such conversational data is informal and occurs in random mix of languages [7]. When this code alternation occurs at or above the utterance level, the phenomenon is referred to as code-switching; when the alternation is utterance internal, the term 'code-mixing'is common [4, 8] constructed social media content code-mixed in three languages Bengali(BN)-English(EN)-Hindi(HI), from Facebook comprising of 2335 posts and 9813 comments.

Annotation at different levels of code-mixing were provided which included sentence-level, fragment-level and word-level tagging. Dictionary based methods of classification were compared with supervised classification methods namely Support Vector Machine (SVM) and sequence labeling using Conditional Random Field (CRF), with and without context information [4], concluded that word-level classifier with contextual clues perform better than unsupervised dictionary based methods. The large volume of code-mixed data on the web has introduced several new challenges and opportunities in tasks like capturing opinions of general public.

Today, sentiment analysis has become a specialized research area finding increasing importance in commercial applications and mining business processes, by virtue of its task of

processing opinion data. In the process, aspect extraction plays a fundamental role which aims at extracting the aspects of an entity in a domain on which opinions have been expressed [11, 15, 22, 25]. Traditionally opinions can be expressed at document level, sentence level and aspect level [15]. To maximize the value from the opinions we need to process opinions at the fine grained level of granularity. So we chose to work at the aspect level. Fundamentally, the task of aspect based opinion analysis comprises of identifying and extracting aspects of an entity in a domain in which opinions have been expressed [15].

With very large volume of noisy code-mixed social media data, we find that useful information is highly dispersed and therefore is to be conveyed through latent semantic aspects. Researchers have successfully used topic models for the purpose of latent aspect extraction and grouping in the form of soft clusters [19, 20, 29]. Probabilistic Latent Semantic Analysis (PLSA) [10] and Latent Dirichlet Allocation (LDA) [5], are recommended unsupervised topic modeling methods for aspect extraction and grouping in multilingual context [30]. Each topic is distribution over words with their associated probabilities and each document is distribution over topics.

[30] put forth that an improved topic representation with language independent word distribution works better on text representations that contain synonymous words. Topic models have showed success in tasks like sentiment analysis [17, 29], word sense disambiguation [13] and modeling similarity of terms [5, 13]. However, all this work has addressed monolingual data and at the most parallel bilingual data. We specifically followed the use of probabilistic topic models in multilingual context since most of the social media content consists of random occurrences of words in code-mixed form.

[23] proposed code-switched LDA (cs-LDA), which is used for topic alignment and to find correlations across languages from code-switched corpora with datasets in English-Chinese and English-Spanish. cs-LDA learns semantically coherent topics over LDA as judged by human annotators. However, content analysis in social media like Twitter pose unique challenges as posts are short and written using multiple languages [26]. They used topic modeling for predicting popular Twitter messages and classifying twitter users and corresponding messages into topical categories. In addition, code-mixed chat data has non-standard terminology which makes the semantic interpretation of aspect clusters challenging.

[3] introduced a method for extracting paraphrases that used bilingual parallel corpora. Using automatic alignment techniques from phrase based statistical machine translation, they show how paraphrase in one language like English can be identified using a phrase in another language as a pivot, conveying the same information. They evaluated the quality of obtained paraphrases and concluded that automatic alignment of context contributes to it.

To address the issue of coherence in topic based aspect extraction, [16] proposed a knowledge based method called Automated Knowledge Learning (AKL), using a three step process: first, LDA was run on the domain corpus, then clustering was performed on the obtained topics and finally in step 3, frequent patterns were mined from the topics in each cluster. Besides human evaluation they used topic coherence to show the improvement in precision over the baselines.

In this paper, we propose a novel LDA based algorithm for clustering code-mixed aspect terms semantically. This method termed lcms-LDA, utilizes information from large multilingual semantic knowledge base called BabelNet [21], as mentioned in [1], for monolingual representation of code-mixed words. The key aspect of our proposed framework is that it leverages augmented synsets across languages, and proposes topic distribution based on semantic similarity between the words. This knowledge is used for clustering and inference in lcms-LDA. As a consequence the proposed method results in automatic retrieval of coherent monolingual clusters from code-mixed input.

## 3 Proposed Model

The structure of our proposed lcms-LDA model is described here. Let V be the vocabulary with $[w_1^L, w_2^L, ..., w_W^L]$, words randomly code-mixed in either of L=$[l_1, l_2, l_3, ..., l_l]$, languages where

$l$ denotes number of languages in which code-mixing has occurred. For each message $m_d^L \in \{1, ...., M_c^L\}$, where $M_c^L$, is the collection of messages code-mixed in L languages. We treat each code-mixed message as a document and generally it refers to at least one aspect. Let K be the number of sets indicating groups of semantically related terms.

Fig. 2, shows graphical representation of our proposed lcms-LDA model. Each circle node indicates a random variable and the shaded node indicates w, which is the only observed variable. $\alpha$ denotes Dirichlet prior for document-topic distribution, which assumes same value for all topics. $\beta$ denotes Dirichlet prior for topic-word distribution. Since the value of $\beta$ determines the number of words per topic, we introduce a random variable $\lambda$, which assigns knowledge for augmentation of semantics to each word.

In code mixed text, input words occur in random mix of different languages due to which semantics is spread across languages. In order to automatically deal with this, $\lambda$ enables augmentation of multilingual synsets as proposed in [1]. This updates values of $\beta$ for each word at each iteration of collapsed Gibbs sampling. Therefore, $\lambda$ guides the clustering process of LDA by mapping probability distribution to semantically stronger groups.

We performed approximate inference in lcms-LDA model using the block Gibbs sampler for the estimation of the posterior distribution of P(z|w ; $\alpha$,$\beta\gamma$). Gibbs sampling computes the conditional distribution to assign topic $z$ and the multilingual synset to a word $w$. Therefore language independent code-mixed topics across the chat collection are given as $Z = [z_1, z_2, z_3, ..., z_k]$.

The conditional distribution of sampling posterior is given in Equation 1:

$$P(z_i, k_i \| z^{-i}, k^{-i}, w, \alpha, \beta, \gamma) \propto \frac{n_{-i}^{z,m} + \alpha}{n_{-i}^{m} + z\alpha} \times \frac{(n^{k,z})_{-i} + \beta}{n_{-i}^{k} + K\beta} \times \frac{(n^{z,k,w_i})_{-i} + \gamma}{n_{-i}^{z,k} + W\gamma}. \quad (1)$$

We have presented the generative process in Algorithm 1. The core aspect behind the proposed lcms-LDA algorithm is that multilingual synset knowledge from the lexical resource adds semantic similarity while word co-occurrence frequencies typically in topic models only grasp syntactic level similarity by statistical means.

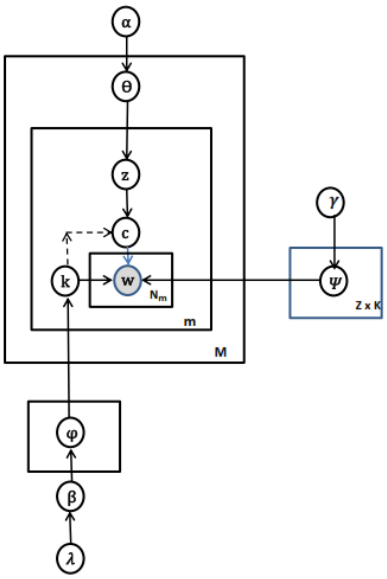


**Fig. 2.** lcms-LDA Plate Notation

Also, since multilingual synsets provide synonyms across languages, monolingual representation of aspects aids in improving the coherence of aspect clusters. Therefore, semantic similarity between words is computed by determining the quantity and quality of overlap across the multilingual synonym lists. We have presented the method in Algorithm 2.

ComputeSemSim $(T_i, T_j)$, method is used to filter semantically similar terms based on the frequency of common synset components using $\left(1 + log\left(\frac{f_c}{f_a}\right)\right)$, where $f_c$ is the summation of count of common synset terms across the words in comparison i.e. $T_i$ and $T_j$ and $f_a$ is the frequency of the referred term i n the context.

## 4 Experimental Results

In this section we first present the implementation details of lcms-LDA and then show the

1. **foreach** *topic* $z \in \{1..Z\}$ **do**
   **foreach** *word w* $\in \{ w_1, ..., w_W \}$ **do**
   $\varphi_{z,i} \sim$ Dir($\beta_i$)
   //distribution of topics over words
   **foreach** *multilingual synset k*
   $\in \{ 1, ..., K \}$ **do**
   $\psi_{z\,x\,k} \sim$ Dir($\gamma$)
   //distribution over multilingual synsets
   **end**
   **end**
   **end**
2. **foreach** *code-mixed message m*$\in \{1...M_c\}$ **do**
   $\theta_m \sim$ Dir($\alpha$)
   //distribution of code-mixed messages over topics
   **foreach** *code-mix word*
   $w_{m,n}$*where language l*$\in$ *L and L =*
   $\{ l_1, l_2, l_{3,...}, l_l \}$ *and n*$\in \{ 1...N_m \}$ **do**
   $z_{m,n} \sim$ Multi($\theta_m$)
   //draw an aspect
   $k_{m,n} \sim \varphi_{z m,n}$
   //draw a multilingual synset
   $w_{m,n} \sim \psi_{z m,n,k m,n}$
   //draw a topic
   **end**
   **end**

**Algorithm 1:** lcms-LDA generative process

1. **foreach** *topic term i of topic cluster* $\{T_1, ..., T_N\}$ **do**
   vec($T_i$)= multiSyn($T_i$)
   // //Obtaining vec($T_i$) from k of lcms-LDA model
   **end**
2. **foreach** $t \in T_i$ **do**
   addAttrib($t, cnt$)
   // //cnt count the frequency of this term across the terms used for comparison
   **end**
3. **foreach** $i \in 1, .., m$ **do**
   **foreach** $j \in i + 1, .., m$ **do**
   sim-score =
   computeSemSim($T_i, T_j$)
   **end**
   **end**

**Algorithm 2:** lcms-Semantic Process

performance of the lcms-LDA aspect extraction framework.

### 4.1 Dataset Used

For the evaluation of the proposed lcms-LDA model we used FIRE 2014[1] (Forum for IR Evaluation), for shared task on transliterated search. This dataset comprises of social media posts in English mixed with six other Indian languages.The English-Hindi corpora from FIRE 2014 was introduced by [7]. It consists of 700 messages with the total of 23,967 words which were taken from Facebook chat group for Indian University students. The data contained 63.33% of tokens in Hindi. The overall code-mixing percentage for English-Hindi corpus was as high as 80% due to the frequent slang used in two languages randomly during the chat [7].

As stated in [29], topic models are applied to documents to generate topics from them. The key step in our method is clustering similar code-mixed words co-occurring in the same context. According to [9], the words occurring in the same context tend to be semantically similar.

The key step in our method is introduction of cluster variable c in lcms-LDA model which groups code-mixed words semantically i.e words co-occurring in the same context. Such words across the languages with similar probabilities belong to the same topic and words with different probabilities are distributed across topics. Since implicitly context is closely shared with a message, we treat each code-mixed message independently. This representation is suitable for us as the word semantic similarity first resolves context at the message level.

$\lambda$ enables augmentation with monolingual synsets for code-mixed words as proposed in [1]. The key step in our method is introduction of cluster variable c in lcms-LDA model which groups monolingual words semantically i.e words co-occurring in the same context. Such words across the languages with similar probabilities belong to the same topic and words with different probabilities are distributed across topics.

In Fig. 3, we demonstrate an example of two topic clusters to show the effect of c. The two sets of words in Fig. 3, show two sample topic clusters formed by top aspects based on the probability of occurrence. The first cluster in Fig. 3, shows a topic cluster with c disabled and the second cluster indicates the aspects clustered with c enabled. In Fig. 3, each aspect is separated by a semi-colon

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

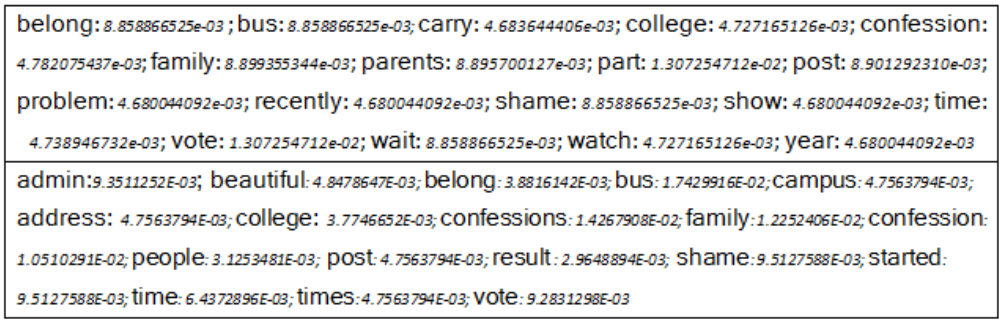


**Fig. 3.** Top sample aspect topic clusters with probability

followed by the respective probability of occurrence shown in italic font.

Fig. 4 shows the sample clusters generated by lcms-LDA.

### 4.2 Data Pre-Processing

In our proposed lcms-LDA model, our objective is to address coherence of aspects, which basically clusters words that are semantically related irrespective of the language in which they are written. At the pre-processing stage we address this need by employing shallow parser [27] and obtain the normalized output. We then used POS tagger by [24], to obtain POS tag of each word. Since our data has random mix of words in different languages, we tagged such words based on the context of the neighboring words. We addressed noise words by eliminating stop words[2] for Hindi and English.

### 4.3 Results

To evaluate the proposed lcms-LDA model on the said dataset, we consider the comparison with the two baselines PLSA [10] and LDA [5]. Also, we tested the performance comparing lcms-LDA with the aspect topic distributions obtained from LDA using augmented monolingual words proposed by [1]. We refer to this output as Aug-Mono-LDA. The language independent code-mixed semantically coherent aspect topics across the chat collection is given as $Z = [z_1, z_2, z_3, ..., z_k]$. For all models, for a single sample, with 200 burn-in iterations, we took posterior inference after 2000 iterations of Gibbs sampling. For other parameter settings we used $\alpha = 1$, $\beta = 0.1$ and $\gamma$ is set to number of words in a cluster. For lcms-LDA we performed testing based on variable number of terms per topic to record its effect on coherence. We tested the efficiency of semantic comparison injected by $\alpha$.

First, lcms-LDA was employed on the dataset described in Section 4.1. Then each message of the said code-mixed chat corpus, after pre-processing was computed as aspect topic distribution based on semantic coherence. It should be noted that work in [1] only described the discovery of language independent aspects and did not include semantics for coherence improvement of aspect clusters. To the best of our knowledge this is the first work addressing extraction of language independent aspects from code-mixed input.

We compare the proposed method with the language independent aspect [1], based LDA. Thus obtained aspect clusters called Aug-Mono clusters indicate augmented monolingual clusters in English or Hindi language of bilingual Hindi-English code-mixed corpus. Their topic distributions are computed on monolingual words based on co-occurrences. We used this method for comparison as we are interested in evaluating coherence by comparing aspect clusters based on semantics against statistically obtained clusters. We evaluate our approach over this system in terms of both aspect cluster interpretibility as well as distinctiveness. Both these measures contribute

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

| LDA | Aug-Mono-Eng | lcms-LDA-Eng | Aug-Mono-Hin | lcms-LDA-Hin |
|---|---|---|---|---|
| life 1.218945711e-02 | of 2.197904038e-02 | ur 1.317345655e-02 | आज 1.326167684e-02 | अधिक 8.637525167E-03 |
| confession 8.179159553e-03 | to 1.871296169e-02 | girl 1.302102844e-02 | के 1.218313916e-02 | ज़िंदगी 7.626218355E-03 |
| लो 6.918641824e-03 | a 1.871296169e-02 | ka 1.106350189e-02 | परमाणु संख्या 7 1.139807026e-02 | मई 7.812915157E-03 |
| साला 5.571348367e-03 | the 1.825382461e-02 | girls 1.089272527e-02 | पैसे 1.139807026e-02 | विद्यार्थी 7.403154056E-03 |
| पास 5.504922277e-03 | in 1.684968713e-02 | time 9.177039556e-03 | इतनी 9.534265822e-03 | आज 1.248162429E-02 |
| लड़की 5.504922277e-03 | is 1.542695190e-02 | apne 8.885078504e-03 | ब 7.801519069e-03 | ऊपर 7.126218143E-03 |
| गयी 5.504922277e-03 | and 1.172543002e-02 | country 8.870949931e-03 | ठाट 7.770690036e-03 | छात्र 7.706232952E-03 |
| coz 5.504922277e-03 | में 8.939443397e-03 | make 8.764075259e-03 | जीवन 7.770690036e-03 | जीवन 7.786218142E-03 |
| प्र 5.504922277e-03 | I 8.936405984e-03 | met 8.764075259e-03 | देश 7.770690036e-03 | ठाट 7.129597392E-03 |
| ब 4.261668769e-03 | it 8.469216485e-03 | friend 8.764075259e-03 | मेईं 7.759043812e-03 | ठाठ 7.706232952E-03 |
| पता 4.251470795e-03 | you 8.464219072e-03 | confession 7.050257542e-03 | एक_ही 7.759043812e-03 | देश 7.763257137E-03 |
| बाट 4.240866171e-03 | this 7.042560920e-03 | ago 6.770260986e-03 | छात्रों 7.670170345e-03 | पैसे 1.339986219E-02 |
| man 4.167345497e-03 | your 6.664088145e-03 | tax 6.766127646e-03 | y 7.670170345e-03 | राष्ट्र 7.763228456E-03 |
| मा 4.167345497e-03 | [ 6.655459458e-03 | aa 6.760788957e-03 | ऊपर 7.670170345e-03 | रुपया 1.312194515E-02 |
| half 4.167345497e-03 | n 6.607523269e-03 | didnt 6.739397001e-03 | न्ही 6.243136290e-03 | वर्तमान 1.411242971E-02 |

**Fig. 4.** Top sample aspect topic clusters with probability

towards evaluation of the quality of topic aspect semantic coherence [18].

The Kullback Leibler (KL) divergence measure [12] offers symmetrical KL-Divergence score for comparing distributions. We compared distinctiveness of the aspects clustered in topic distributions produced by [2] and we referred to the same as Code-Mixed PLSA (CM-PLSA). We computed KL-Divergence score for all symmetric combinations and averaged it across all the clusters in a topic. We recorded the scores for different values of $z$ as presented in Fig. 5.

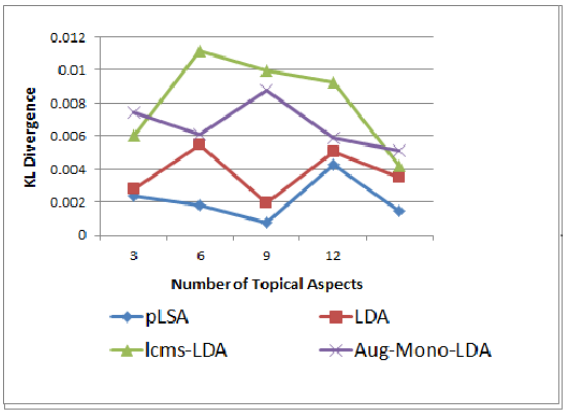


**Fig. 5.** Topic distinctivity comparison for lcms-LDA

From Fig. 5, we observe that the KL-divergence for lcms-LDA is maximum at $z = 6$, as the semantic dissimilarity is maximum, and the overlap results in minimum sore. Also, it is minimum at $z = 3$ due to lesser number of terms generating minimal overlap.

However, as compared to all the models lcms-LDA generates higher distinctiveness and therefore semantics helps in improvement of coherence of aspects resulting in better topic association. The drop in distinctiveness for higher values of $z$ is due to the semantics having high dispersion in chat.

We evaluate the coherence of aspect clusters by yet another standard measure called UMass score. The UMass coherence score is computed by pairwise score function measuring empirical probability of common words [18].

Fig. 6 shows testing for topic interpretibility of topics in each topic cluster. It is maximum at $z = 9$, as the probability of overlap across semantic words in augmentation of each word is maximum, thereby grouping them together. For higher values of $z$, we see the drop in the coherence.

On the semantic evaluation of the words participating in a cluster, we observed that the chat data resulted in inclusion of many ungrammatical words and words which were not nouns. Therefore, with increasing size of the cluster the topic interpretibility is observed to have decreased.

## 5 Conclusion

Active interaction of people on the web through social networks and online text forums is increasingly becoming popular as it encourages communication in random mix of languages. As a consequence,

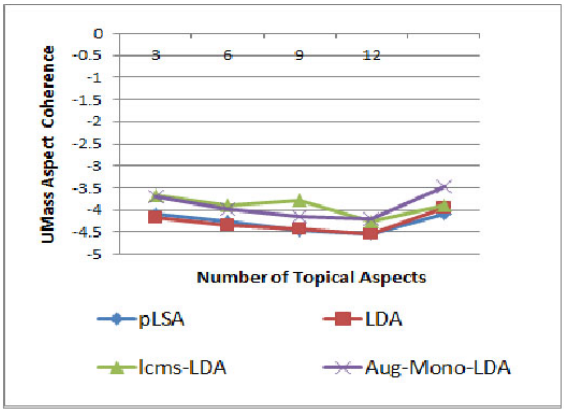


**Fig. 6.** Topic interpretibility comparison for lcms-LDA

tremendous amount of code-mixed data having interesting patterns of interest hidden in it is generated. Sentiment analysis tools generally working on opinion data is always hungry for extraction of useful aspects which are indicators of sentiments and opinions that are implicitly expressed.

In this work, we presented a novel model, termed lcms-LDA, which automatically integrates semantics in the computation of language independent representations of code-mixed words in the LDA algorithm, thus enabling semantics in code-mixed aspect extraction.

Thus, proposed language independent semantic approach to code-mixed aspect extraction, leverages knowledge from freely available in external multilingual resource, thereby introducing automatic clustering of coherent aspect clusters. Therefore, in the perspective of its application, this could be a very useful aid for code-mixed aspect based sentiment analysis.

## References

1. **Asnani, K. & Pawar, J. (2016).** Discovering language independent latent aspect clusters from code-mixed social media text. *Proceedings of the 30th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, Association for the Advancement of Artificial Intelligence(AAAI), pp. 592–595.
2. **Asnani, K. & Pawar, J. (2016).** Discovering thematic knowledge from code-mixed chat messages using topic model. *Third WILDRE Proceedings of the 3rd WILDRE, (LREC)*, pp. 104–109.
3. **Bannard, C. & Callison-Burch, C. (2005).** Paraphrasing with bilingual parallel corpora. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 597–604.
4. **Barman, U., Das, A., Wagner, J., & Foster, J. (2014).** Code mixing: A challenge for language identification in the language of social media. *EMNLP*, volume 13.
5. **Blei, D., Ng, A., & Jordan, M. (2003).** Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, pp. 993–1022.
6. **Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009).** Reading tea leaves: How humans interpret topic models. *Nips*, volume 31, pp. 1–9.
7. **Das, A. & Gambäck, B. (2014).** Identifying languages at the word level in code-mixed indian social media text. *11th International Conference on Natural Language Processing (ICON)*, International Institute of Information Technology.
8. **Gambäck, B. & Das, A. (2016).** Comparing the level of code-switching in corpora. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pp. 1850–1855.
9. **Heinrich, G. (2009).** A generic approach to topic models. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 517–532.
10. **Hofmann, T. (1999).** Probabilistic latent semantic analysis. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., pp. 289–296.
11. **Hu, M. & Liu, B. (2004).** Mining and summarizing customer reviews. *Proceedings of the tenth International conference on Knowledge discovery and data mining (ACM SIGKDD)*, pp. 168–177.
12. **Johnson, D. & Sinanovic, S. (2001).** Symmetrizing the kullback-leibler distance. *IEEE Transactions on Information Theory*.
13. **Lacoste-Julien, S., Sha, F., & Jordan, M. (2009).** Disclda: Discriminative learning for dimensionality reduction and classification. *Advances in neural information processing systems*, pp. 897–904.

14. **Ling, W., Xiang, G., Dyer, C., Black, A. W., & Trancoso, I. (2013).** Microblogs as parallel corpora. *ACL (1)*, pp. 176–186.
15. **Liu, B. (2012).** Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, Vol. 5, No. 1, pp. 1–167.
16. **Liu, Z., Chen, A., & Mukherjee, B. (2014).** Aspect extraction with automated prior knowledge learning. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.
17. **Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007).** Topic sentiment mixture: modeling facets and opinions in weblogs. *Proceedings of the 16th international conference on World Wide Web*, ACM, pp. 171–180.
18. **Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011).** Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 262–272.
19. **Moghaddam, S. & Ester, M. (2013).** The flda model for aspect-based opinion mining: Addressing the cold start problem. *Proceedings of the 22nd international conference on World Wide Web*, ACM, pp. 909–918.
20. **Mukherjee, A. & Liu, B. (2012).** Aspect extraction through semi-supervised modeling. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, Association for Computational Linguistics, pp. 339–348.
21. **Navigli, R. & Ponzetto, S. (2012).** Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, Vol. 193, pp. 217–250.
22. **Papadimitriou, C. & Steiglitz, K. (1982).** *Combinatorial optimization: algorithms and complexity*. Courier Corporation.
23. **Peng, N., Wang, Y., & Dredze, M. (2014).** Learning polylingual topic models from code-switched social media documents. *ACL*, pp. 674–679.
24. **Petrov, S., Das, D., & McDonald, R. (2011).** A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.
25. **Popescu, A. & Etzioni, O. (2005).** Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, Springer, pp. 339–346.
26. **Ramage, D., Dumais, S. T., & Liebling, D. (2010).** Characterizing microblogs with topic models. *ICWSM*, Vol. 10, pp. 130–137.
27. **Sharma, A., Gupta, S., Motlani, R., Bansal, P., Srivastava, M., Mamidi, R., & Sharma, D. M. (2016).** Shallow parsing pipeline for hindi-english code-mixed social media text. *arXiv preprint arXiv:1604.03136*.
28. **Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., et al. (2014).** Overview for the first shared task on language identification in code-switched data. *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pp. 62–72.
29. **Titov, I. & McDonald, R. T. (2008).** A joint model of text and aspect ratings for sentiment summarization. *ACL*, volume 8, pp. 308–316.
30. **Vulić, I., De Smet, W., Tang, J., & Moens, M.-F. (2015).** Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, Vol. 51, No. 1, pp. 111–147.


*Article received on 10/08/2016; accepted on 14/10/2016.*
*Corresponding author is Kavita Asnani.*