Use of Semantic Knowledge Base for Enhancement of Coherence of Code-mixed Topic-Based Aspect Clusters

Kavita Asnani
Computer Engineering Dept
Goa College of Engineering
Goa
India
kavita@gec.ac.in

Jyoti D Pawar
Computer Science and Technology Dept
Goa University
Goa
India
jyotidpawar@gmail.com

Abstract

In social media code-mixing is getting very popular due to which there is enormous generation of noisy and sparse multilingual text which exhibits high dispersion of useful topics which people discuss. Also, the semantics is expressed across random occurrence of code-mixed words. In this paper, we propose **c**ode-**m**ixed **k**nowledge based LDA (cmk-LDA), which infers latent topic based aspects from code-mixed social media data. We experimented on FIRE 2014, a code-mixed corpus and showed that with the help of semantic knowledge from multilingual external knowledge base, cmk-LDA learns coherent topic-based aspects across languages and improves topic interpretibility and topic distinctiveness better than the baseline models . The same is shown to have agreed with human judgment.

1 Introduction

The huge amount of social media text available online is becoming increasingly popular thereby providing an additional opportunity of mining useful information from it. Therefore, most of the research on social media text has concentrated on English chat data or on multilingual data where each message as a component is monolingual. In social media, people often switch between two or more languages, both at conversation level and at message level (Ling et al., 2013). However, majority of conversational data on social networking forums is informal and occurs in random mix of languages (Das and Gambäck, 2014). 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 (Gambäck and Das, 2016). Thus, code-mixing while chatting has become prevalent in current times. However, exponentially increasing large volumes of short and long code-mixed messages contain lot of noise and has useful information highly dispersed. Unfortunately, it is not an easy task to retrieve useful knowledge from such data as code-mixing occurs at different levels of code-complexity and imposes fundamental challenges namely:

1. Code-mixed social media data is multilingual, usually bilingual (San, 2009). Therefore, semantics is spread across languages.

2. Social media data do not have specific terminology (Eisenstein, 2013).

Therefore, using training data from parallel or comparable corpora will not be useful in this context. Also, availability of pre annotated corpora for all language pairs used in social media may practically be very difficult to obtain. Our objective is to model unsupervised aspect extraction using topics, from the code-mixed context to obtain useful knowledge. Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA)(Blei et al., 2003) are popularly recommended unsupervised topic modeling methods for this purpose; but a shortcoming with them is that they result in extracting some incoherent topics (Chang et al., 2009).This is due to the occurrence of some irrelevant and polysemous terms in extraction (Chang et al., 2009), which is likely to get aggregated in multilingual content. Therefore, in code-mixed context extraction of incoherent topics are highly likely to occur. In order to cross over the word level language barrier and augment each word with its semantic description we propose to leverage knowledge from external multilingual semantic knowledge base such as BabelNet v3.0 [1] (Navigli et al., 2012) which is created

[1] http://babelnet.org

D S Sharma, R Sangal and A K Singh. Proc. of the 13th Intl. Conference on Natural Language Processing, pages 259–266,
Varanasi, India. December 2016. ©2016 NLP Association of India (NLPAI)

from integration of both Wikipedia and WordNet (Miller, 1995).

Our approach of incorporating semantic knowledge in LDA topic model has resemblance with General Knowledge based LDA (GK-LDA)(Chen et al., 2013) model. However, their sets were monolingual and were constructed using WordNet. They called their semantic sets as Lexical Relation Sets(LR-sets) and were comprising of synonym, antonym and adjective relations. They addressed the problem of multiple senses using synonyms and antonyms in LR-Sets. However, in their work since semantic knowledge is augmented at word level not all LR-sets resulted in correct context knowledge. They handled this problem by using explicit word-correlation matrix and also fixed the wrong knowledge explicitly. But we take a different approach. In a single step, we obtain disambiguated synsets for a code-mixed message, retrieving correct multilingual synsets appropriate to the context across languages. This also resolves language related multiple senses issue. As per our knowledge, our proposed model is the first model to exploit semantic knowledge from BabelNet in topic models for producing coherent topics from code-mixed data.

The remaining part of the paper is organized as follows: Section 2 presents related work, Section 3 describes our proposed work, Section 4 gives implementation details, Section 5 presents experimental results and Section 6 states the conclusion.

2 Related Work

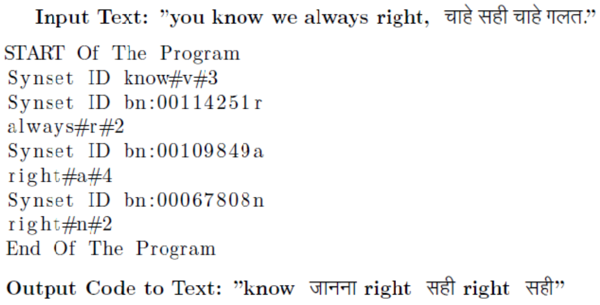
Code-switching and code-mixing are popularly been observed on social networking forums, especially in highly multilingual societies. In this paper, we will use the term code-mixing to refer to both of these situations. In general, multilingual users communicate in two or more languages. In (Barman et al., 2014) study was performed on code-mixed content by collecting data from Facebook posts, code-mixed in three languages Bengali(BN)-English(EN)-Hindi(HI). However, in (Bali et al., 2014), the analysis of code-mixed bilingual English-Hindi data showed significant amount of code-mixing and proved that most of the active users on facebook are bilingual. Out of 43,012 words on facebook chat, 38,144 were written in Roman script and 2661 in Devanagari script. They claimed that the deeper analysis of such data requires discourse linguists. We first briefly describe survey studies addressing characteristics of code-mixed text content which are affected by linguistic challenges. Linguistic processing of code-switched data was done by (Vyas et al., 2014) (Sequiera et al., 2015)(Solorio and Liu, 2008)(San, 2015) and they concluded that natural language processing tools are not equipped to resolve all issues related to pre-processing of code-mixed text. (Vyas et al., 2014) created multi-level annotated corpus of Hindi-English code-mixed text from Facebook forums. The annotations were attributed across three levels representing POS tags, language identification and transliteration respectively. They proved that POS tagging of code-mixed text cannot be done by placing two or more monolingual POS taggers together. They identified that it is complex to deal with code-mixed data due to spelling variation. (Sequiera et al., 2015) experimented with joint modeling of language identification and POS tagging. (San, 2015) used Random Forest based method on 400 code-mixed utterances from Facebook and Twitter and reported 63.5% word level tagging accuracy. Their work also illustrates challenge in POS tagging and need for transliteration.

Thus, code-mixed social media data suffer from its associated linguistic complexities which make the semantic interpretation of such high dimensional content very challenging. Also, content analysis in social media containing Twitter posts where training and evaluation is concerned, is done using supervised models in machine learning and NLP (Ramage et al., 2010). They claimed that this requires good prior knowledge about data. Such supervised approaches are not feasible in the context of code-mixed chat data as such data is generated randomly in any language and availability of parallel or comparable corpora for training in certain language pairs is difficult. Also, the use of machine translators is very challenging and not feasible in social media context as the volume of the data is large, inconsistent and translation is required to be done at the word-level. Using a shallow parser (Sharma et al., 2016), our proposed approach first addresses noise elimination and need for normalization. Then, we were motivated to model appropriate unsupervised topic based aspect extraction to discover useful knowledge offering significant information to the administrator or end user. Therefore, we turned to unsupervised topic models, as our goal is to use large collection of code-mixed documents and convert it into aspects of the text in the form of clusters of similar themes together called as topics.

(Peng et al., 2014) proposed code-switched LDA (cs-LDA) which is used for topic alignment and to find correlations across languages from code-switched social media text. They used two code-switched corpora (English- Spanish Twitter data and English-Chinese Weibo data) and performed per topic language-specific word distribution and per word language identification. They showed that cs-LDA improves perplexity over LDA, and learns semantically coherent aligned topics as judged by the human annotators. In addition, as code-mixed social media data involve words occurring in random mix of languages, very few languages have word-level language identification systems in place.

In this paper, we propose utilizing information from large multilingual semantic knowledge base called BabelNet (Navigli et al., 2012) as it provides the same concept expressed in many different languages; which can be used to augment information to code-mixed words, thereby dropping the language barrier. BabelNet (Navigli et al., 2012) offers wide coverage of lexicographic and encyclopedic terms. BabelNet provides multilingual synsets where each synset is represented by corresponding synonymous concepts in different languages. BabelNet v3.0 offers coverage for 271 languages with 14 million entries comprising of 6M concepts, 745M word senses and 380M semantic relations.The knowledge incorporated from BabelNet is used to guide our proposed cmk-LDA topic model.

In order to handle wrong knowledge (injected due to multiple senses), Babelfy (Moro et al., 2014) is used to obtain disambiguated code-mixed words across the message. Babelfy leverages information from BabelNet for its joint approach to multilingual word sense disambiguation and entity linking. It performs semantic interpretation of an ambiguous sentence using a graph and then extracts the densest subgraph as the most coherent interpretation. In order to discover knowledge from social media (Manchanda, 2015) investigated to find new entities and disambiguation as a joint task on short microblog text. They aimed to improve the disambiguation of entities using linked datasets and also discovery of new entities from tweets, thus improving the overall accuracy of the system.



```
Input Text: ”you know we always right, चाहे सही चाहे गलत.”
START Of The Program
 Synset ID know#v#3
 Synset ID bn:00114251r
 always#r#2
 Synset ID bn:00109849a
 right#a#4
 Synset ID bn:00067808n
 right#n#2
 End Of The Program
 Output Code to Text: ”know जानना right सही right सही”
```

Figure 1: An Example

3 Our Proposed Work

In this section we introduce our proposed work. We first present how we addressed random occurrence of words in different languages in a code-mixed message. For this purpose, we utilize knowledge from BabelNet which helps augment semantic interpretations of code-mixed words across languages in the form of **m**ultilingual **k**nowledge **S**ets (mulkSets). We propose a new knowledge based LDA for code-mixed data, which we have called **c**ode-**m**ixed **k**nowledge based LDA (cmk-LDA). In order to automatically deal with random words occurring in different languages we add a new latent variable k in LDA, which denotes the mulkSet assignment to each word. We initially tried to construct mulkSets by directly obtaining synsets from BabelNet. But for each code-mixed word it resulted in retrieval of large number of synsets. This is due to all the possible multilingual senses assigned at the word level. Therefore, for correct semantic interpretation we had to ensure that code-mixed words sharing the same context should share similar sense in their mulkSets. Hence, we constructed mulkSet with disambiguated synsets using BabelFy [2]. Such disambiguated synsets address the shared context across languages. An example is presented in Figure 1. Our disambiguated set therefore contains the revised vocabulary having three words. We found this knowledge to be beneficial to our proposed cmk-LDA topic model as each topic is a multinomial distribution over mulkSets. Thus, cmk-LDA model finds co-occuring words automatically in a language independent manner. For the purpose of demonstration at the sentence level we illustrate two instances in the Figure 3. The

[2] http://babelfy.org

first sentence in Figure 3 is an input code-mixed sentence to Babelfy and the second sentence is the disambiguated output further augmented with synsets across languages from BabelNet. Based on this knowledge, cmk-LDA model further generates topic-based aspects across sentences by processing probability distribution over mulkSets.

3.1 The cmk-LDA Model

Given a collection of code-mixed messages
M= { $m_1^L, m_2^L, m_3^L, ..., m_n^L$ }
where n denotes number of code-mixed messages in L=$l_1, l_2, l_{3,...}, l_l$ languages where l denotes number of languages in which code-mixing has occurred.

The code-mixed message is represented as

$$m_i^L = \{w_{i1}^L, w_{i2}^L\, w_{i3}^L \ldots, w_{iN_i}^L\}$$

where N_i denotes number of words in the i^{th} message and w_{ij} denotes j^{th} word of the i^{th} message.

Figure 2 shows graphical representation for our proposed cmk-LDA model. Each circle node indicates a random variable and the shaded node indicates w, which is the only observed variable. We introduce new latent variable k, which assigns mulkSet to each word. Assume that there are K mulksets in total. Therefore, language independent code-mixed topics across the chat collection are given as: Z= { $z_1, z_2, z_{3,...}, z_k$ }

Each code-mixed message is thus considered as a mixture of K latent code-mixed topics from the set Z. These topic distributions are modeled by probability scores $P(\ z_k \mid m^i)$ Thus, M is represented as set Z of latent concepts present in M.

We have presented the generative process in Algorithm 1.

We performed approximate inference in cmk-LDA model using the block Gibbs sampler as followed typically in LDA. Gibbs sampling constructs Markov chain over latent variables, by computing the conditional distribution to assign a topic z and the mulkSet k to the word. The conditional distribution for sampling posterior is given in Equation 1.

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

```
1. foreach topic $z \in Z$ do
       Draw mulkSet distribution
       $\varphi \sim$ Dir($_\beta$)
       foreach mulkSet $k \in \{ 1, ..., K \}$ do
           Draw mulkSet distribution over words
           $\psi_{z\,x\,k} \sim$ Dir($_\gamma$)
       end
   end

2. foreach code-mixed message $m \in M$ do
       Draw topic distribution $\theta_m \sim$ Dir($_\alpha$)
       foreach code-mix word
       $w_{m,n}^{l}$ where language $l \in L$ and $L = \{ l_1, l_2, l_{3,...}, l_l \}$ and $n \in \{ 1 ... N_m \}$ do
           Draw a topic $z_{m,n} \sim \theta_m$
           Draw a k-mulkSet $k_{m,n} \sim \varphi_{zm,n}$
           Draw a topic $w_{m,n} \sim \psi_{\ zm,n,km,n}$
       end
   end
```

Algorithm 1: cmk-LDA Generative Process

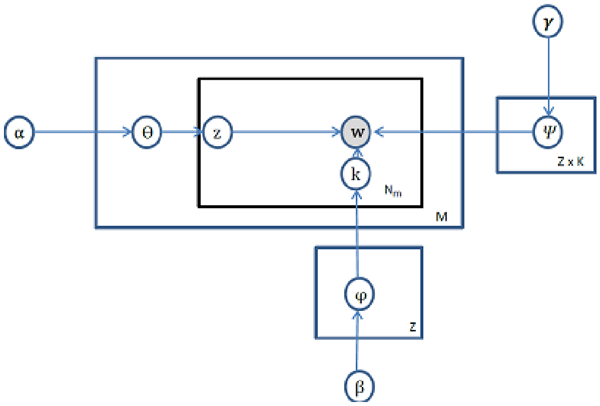
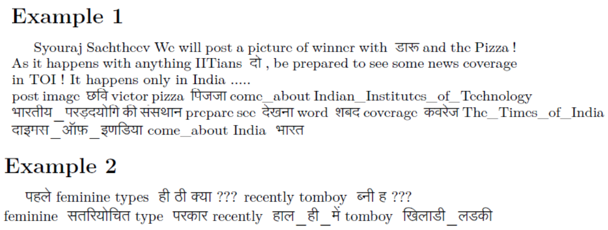


Figure 2: cmk-LDA Plate Notation



Example 1

Syouraj Sachtheev We will post a picture of winner with डारू and the Pizza !
As it happens with anything IITians दो , be prepared to see some news coverage
in TOI ! It happens only in India .....
post image छवि victor pizza पिजजा come_about Indian_Institutes_of_Technology
भारतीय_परड़दयोगि की संसथान prepare see देखना word शबद coverage कवरेज The_Times_of_India
दाइमस_ऑफ़_इणडिया come_about India भारत

Example 2

पहले feminine types ही ठी क्या ??? recently tomboy ब्नी ह ???
feminine सतरियोचित type परकार recently हाल_ही_में tomboy खिलाडी_लडकी

Figure 3: Example Sentences

Figure 6 shows the sample clusters generated by cmk-LDA.

4 Implementation Details

We have evaluated the proposed cmk-LDA model and compare it with the two baselines pLSA (Hofmann, 1999) and LDA(Blei et al., 2003). In our experiments, our proposed cmk-LDA model can deal with words randomly sharing context in either of the two languages Hindi or English. We compared the models by measuring coherence of aspect clusters. For evaluating semantic coherence of topics we used two evaluation metrics; topic coherence(UMass) and KL-Divergence to measure topic interpretability and topic distinctiveness respectively. We performed experiments on four models based on use of external semantic knowledge. The two baselines are addressed as *wek-PLSA* and *wek-LDA* where wek indicates models **w**ithout **e**xternal **k**nowledge. We perform testing with different number of topics k and we made sure that we compare topic aspect clusters of the same size.

4.1 Dataset Used

We performed experiments on FIRE 2014[3](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 (Das and Gambäck, 2014). 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 (Das and Gambäck, 2014).

4.2 Code-mixed data pre-processing

In our proposed cmk-LDA topic model we believe that a topic is semantically coherent if it assigns high probability scores to words that are semantically related irrespective of the language in which they are written. In the pre-processing phase, we used Shallow parser (Sharma et al., 2016) for our purpose to obtain normalized output. (Sharma et al., 2016) experimented on the same code-mixed English-Hindi FIRE 2014 dataset as we did and they reported the accuracy on parsing as 74.48% and 75.07% respectively. Basically noise was eliminated by removal of stop-words [4] for Hindi and English.

Table 1: Cohens Kappa for inter-rater agreement

Index	1	2	3	4	5
k	3	6	9	12	15
Precision@k	0.899	0.798	0.876	0.712	0.766

4.3 Code-Mixed Message as a Document

Topic models are applied to documents to produce topics from them (Titov and McDonald, 2008). The key step in our method is to determine context and for that we address the code-mixed words cooccurring in same context. Such words with similar probabilities belong to the same topic and rejects words that have different probabilities across topics. Therefore, we treat each code-mixed message independently. Although, relationship between messages is lost, the code-mixed words across the language vocabulary within the code-mix message contribute to the message context. This representation is fair enough as it is suitable to obtain relevant disambiguated sets from Babelfy as it resolves context at the message level.

5 Experimental Results

5.1 Measuring Topic Quality by Human Judgement

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 in the steps of 5, 10, 15 and 20 code- mixed aspects generated by cmk-LDA model and sorted in the descending order of probabilities under that topic. For each topic, the judges annotated the topics at word level and then we aggregated their results to annotate the cluster. Table 1 reports the Cohens Kappa score for topic annotation, which is above 0.5, indicating good agreement. We observed a high score at k=3 due to few aspect topics with context highly dispersed resulting in strong agreement on low quality clusters. According to the scale the Kappa score

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

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

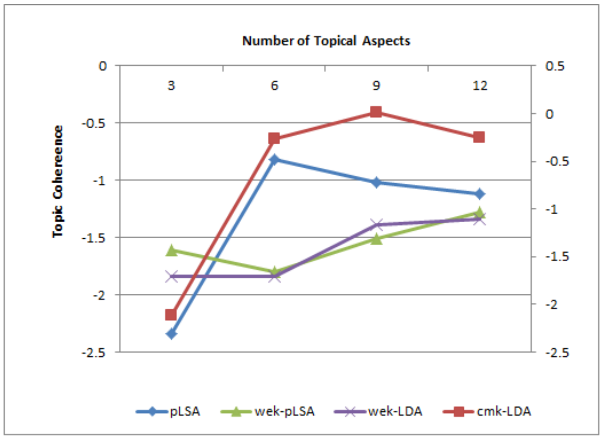


Figure 4: Topic Interpretibility Comparison

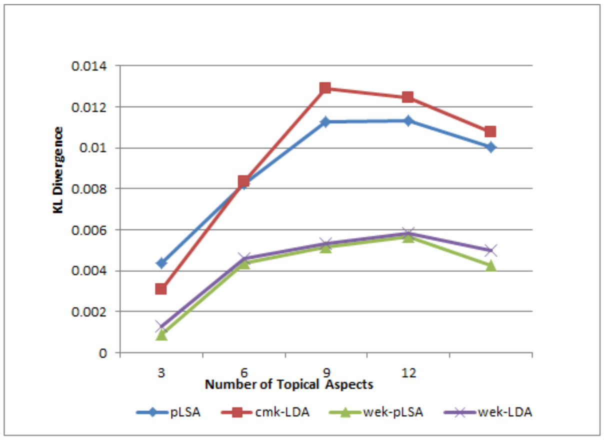


Figure 5: Topic Distinctivity Comparison

increases with more number of topic-based aspect clusters as the topics get semantically stronger. Relatively high agreement at k=9 points to the likely generation of good quality aspect topics.

5.2 Measuring Topic Interpretability

The UMass measure uses a pairwise score function introduced by (Mimno et al., 2011). The score function is not symmetric as it is an increasing function of the empirical probability of common words. Figure 4 shows testing for topic interpretability of topic based aspect clusters for comparison of all models with and without external knowledge. We see from the trend generated by cmk-LDA relative to the other models, generates rise indicating enhancement in coherence of topic distributions. Since most of the topics across languages indicate common context, such high probability topics in a cluster seem to be contributing to high UMass coherence. Both pLSA and cmk-LDA offer better topic interpretibility of clusters. These results confirm that incorporating multilingual external semantic knowledge in code-mixed data find higher quality topics offering better interpretation.

5.3 Measuring Topic Distinctiveness

The Kullback Leibler (KL) divergence measure (Johnson and Sinanovic, 2001) is a standard measure for comparing distributions. We apply symmetrical version of KL Divergence and average it across the topics. Higher KL divergence score indicates that the words across the topics are distinct and are considered to generate higher topic distinctiveness. From the Figure 5, we can see that initially at k=3, cmk-LDA is lower because the context is wide and therefore words are common across the clusters. While KL scores of pLSA model closely follow cmk-LDA, topics at k=9 have generated higher KL score for cmk-LDA. This suggests that though the interpretibility of topics gains is as high in cmk-LDA, the topics are more distinct as well.

6 Conclusion and Future Work

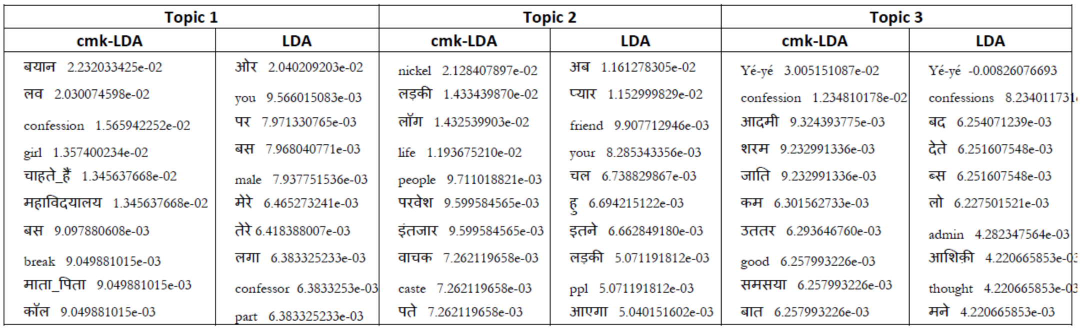
In order to enhance coherence of topic-based aspects discovered from code -mixed social media data, we proposed a novel unsupervised cmk-LDA topic model which utilizes semantic knowledge from external resources. Such knowledge resolves semantic interpretations across random mix of languages. Our evaluation results show that cmk-LDA outperforms the baselines. We state that our proposed framework supports the utility of lexical and semantic knowledge freely available in external multilingual resources which can drop the language barrier and can help discover useful aspects.

We have not addressed mixed-script in our experiments. In our future experiments we will explore methods to perform term matching and spelling variation modelling the terms across the scripts.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Jonathan Chang, Sean Gerrish, Chong Wang, Jor-



Topic 1		Topic 2		Topic 3	
cmk-LDA	LDA	cmk-LDA	LDA	cmk-LDA	LDA
बयान 2.232033425e-02	ओर 2.040209203e-02	nickel 2.128407897e-02	अब 1.161278305e-02	Yé-yé 3.005151087e-02	Yé-yé -0.00826076693
लव 2.030074598e-02	you 9.566015083e-03	लड़की 1.433439870e-02	प्यार 1.152999829e-02	confession 1.234810178e-02	confessions 8.23401173
confession 1.565942252e-02	पर 7.971330765e-03	लॉग 1.432539903e-02	friend 9.907712946e-03	आदमी 9.324393775e-03	बद 6.254071239e-03
girl 1.357400234e-02	बस 7.968040771e-03	life 1.193675210e-02	your 8.285343356e-03	शरम 9.232991336e-03	देते 6.251607548e-03
चाहते_हैं 1.345637668e-02	male 7.937751536e-03	people 9.711018821e-03	चल 6.738829867e-03	जाति 9.232991336e-03	ब्स 6.251607548e-03
महाविदयालय 1.345637668e-02	मेरे 6.465273241e-03	परवेश 9.599584565e-03	हु 6.694215122e-03	कम 6.301562733e-03	लो 6.227501521e-03
बस 9.097880608e-03	तेरे 6.418388007e-03	इंतजार 9.599584565e-03	इतने 6.662849180e-03	उततर 6.293646760e-03	admin 4.282347564e-03
break 9.049881015e-03	लगा 6.383325233e-03	वाचक 7.262119658e-03	लड़की 5.071191812e-03	good 6.257993226e-03	आशिक़ी 4.220665853e-0
माता_पिता 9.049881015e-03	confessor 6.3833253e-03	caste 7.262119658e-03	ppl 5.071191812e-03	समसया 6.257993226e-03	thought 4.220665853e-0
कॉल 9.049881015e-03	part 6.383325233e-03	पते 7.262119658e-03	आएगा 5.040151602e-03	बात 6.257993226e-03	मने 4.220665853e-03

Figure 6: Sample Topics with Probability (Top n probability aspects comprise topics)

dan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

Amitava Das and Björn Gambäck. Identifying languages at the word level in code-mixed indian social media text. 2014.

Jacob Eisenstein. What to do about bad language on the internet. In *HLT-NAACL*, pages 359–369, 2013.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

Wang Ling, Guang Xiang, Chris Dyer, Alan W Black, and Isabel Trancoso. Microblogs as parallel corpora. In *ACL (1)*, pages 176–186, 2013.

Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

Hong Ka San. Chinese-english code-switching in blogs by macao young people. *Master's thesis, The University of Edinburgh, UK*, 2009.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. Code mixing: A challenge for language identification in the language of social media. *EMNLP 2014*, 13, 2014.

Anupam Jamatia, Björn Gambäck, and Amitava Das. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. *Recent Advances in Natural Language Processing, Bulgaria*, page 239–248, 2015.

Nanyun Peng, Yiming Wang, and Mark Dredze. Learning polylingual topic models from code-switched social media documents. In *ACL (2)*, pages 674–679, 2014.

Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. *ICWSM*, 10:1–1, 2010.

Royal Sequiera, Monojit Choudhury, and Kalika Bali. Pos tagging of hindi-english code mixed text from social media: Some machine learning experiments. *ICON 2015*

Kalika Bali Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. i am borrowing ya mixing? an analysis of english-hindi code mixing in facebook. *EMNLP 2014*, page 116, 2014.

Thamar Solorio and Yang Liu. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060. Association for Computational Linguistics, 2008.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. Pos tagging of english-hindi code-mixed social media content. In *EMNLP*, volume 14, pages 974–979, 2014.

Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Srivastava, Radhika Mamidi, and Dipti M Sharma. Shallow parsing pipeline for hindi-english code-mixed social media text. *arXiv preprint arXiv:1604.03136*, 2016.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244, 2014.

Björn Gambäck, and Amitava Das. Comparing the Level of Code-Switching in Corpora. *LREC*, 2016.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 209–218. ACM, 2013.

Jason Chuang, Sonal Gupta, Christopher D Manning, and Jeffrey Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *ICML (3)*, pages 612–620, 2013.

Don Johnson and Sinan Sinanovic. Symmetrizing the kullback-leibler distance. *IEEE Transactions on Information Theory*, 2001.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Pikakshi Manchanda. Entity linking and knowledge discovery in microblogs. *ISWC-DC 2015 The ISWC 2015 Doctoral Consortium*, page 25, 2015.

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM, 2008.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.