J. Inst. Eng. India Ser. B (April 2022) 103(2):591–597
https://doi.org/10.1007/s40031-021-00655-w

ORIGINAL CONTRIBUTION

# Proposed Model for Context Topic Identification of English and Hindi News Article Through LDA Approach with NLP Technique

**Anukriti Srivastav**[1] · **Satwinder Singh**[1]

Received: 4 June 2020 / Accepted: 12 July 2021 / Published online: 14 August 2021


**Abstract** According to the survey, India has the world's second-largest newspaper market, with more than 100 K newspaper outlets, approx 240 million circulation, and 1300 million subscribers or readers. The topic modeling work is increasing day by day, and researchers have published multiple topic modeling papers and have implemented them in different areas like software engineering, political science and medical, etc. LDA topic modeling is used in this research because it has been introduced successfully for topic modeling and classification and it measures the probability of a text-dependent on the bag-of-words scheme without considering the word series. LDA is a common topic modeling algorithm with excellent implementation in the Gensim Python package. However, the challenge is how to extract good quality topics that are simple, separated, and meaningful. The purpose of this research deals with finding the main topics of the same category news articles which are in two different languages (Hindi and English) and then classifying these different language news topics with similarity measurement. In this research, the corpus is constructed with bigram. To achieve the research goal, we have to first build a headline and link extractor that scrap the top news from Google News feeds for both English and Hindi languages (Google News collects news stories that have appeared on different news website which is already accessible in 35 languages over the last 30 days) and then analyses which two news headlines are similar.

✉ Anukriti Srivastav
anukritisrivastava072@gmail.com

[1] Centre for Computer Science and Technology, Central University of Punjab, Bathinda, India

**Keywords** Text mining · Topic modeling using LDA · Natural language processing (NLP) · Similarity measurement with LDA cosine similarity

## Introduction

The quantity of valuable data produced by humans has caused competition over the last decade. Nevertheless, according to recent research, approximately 2.6 quintillion bytes of data are generated every day, and 90% of the world's data have been generated in the previous four to three years [1]. When the quantity of data obtains grows exponentially, the valuable data that online users are searching for become difficult to obtain and process. Topic modeling is a widely used data extraction method to discover the latent semantics structure in a frame of text which offers an excellent approach to evaluate broad undefined documents. The topic modeling in artificial intelligence and natural language processing is a form of a mathematical algorithm for finding the specific "topics" which appear in a set of documents. It is also known as mixed-membership models [2].

Any topic includes a series of many words which often overlap. The modeling approach to the topic may link words with related contexts and also identify dual-significants word uses.

Many topic models consider that texts are monolingual; however, some may capture statistical dependencies among multiple representation groups that may be used for parallel or comparable multilingual texts [3].

In this research, a comparable text is a combined text consisting of two different languages (English and Hindi) parts in which Hindi is the translation of the English

language, often having alignments between a word to word or sentence to sentence.

English and Hindi are one of the world's most common languages. There are several English and Hindi news article websites. LDA has done many research works for the categorization of undefined texts. The latest of these works is a generative LDA model designed to identify texts in which are in the English language [4]. Every news article's document for this research work contains a mixture of topics. Throughout these contexts, an LDA model is build to find topics from a news article corpus that are in English and Hindi language. The researcher evaluated a perplexed topic model [4]. Perplexity has historically been used several times as a method of evaluation for the derived topics, but it has been found that it often does not correspond with human annotations [5]. A systematic approach to text mining was suggested in [6]. The research has interacted with Wikipedia and Twitter data here. For those two datasets, two different perspectives were discussed. A document topic model was achieved from the Wikipedia data, aiming at a topic-wise content search [6]. On the other hand, a user topic model was explored with the Twitter data to classify the interest of users in Twitter data. This concept can also be put into action on two different languages newspaper data to analyze news patterns over a certain period. Also, similarity calculation for the two different languages (English and Hindi) news article data was measured in this paper. As all know, several languages have various grammatical structures and also stemming techniques so it is difficult to get core topics from several languages.

The goal of this research is to find the core topics of (approx) same category news which are in two different languages, i.e, English and Hindi and then also measure the similarity score between these two different languages news articles topics. To achieve the research goal, this work has to build a headline and link extractor that scrap the top news from Google News feeds for both English and Hindi languages and then analyses which two news headlines are similar. After collecting the news articles firstly, we have to translate the Hindi news article into English by using the "Google Translator" Python library. The basic idea to find the core topics from LDA of the same category news is to calculate the coherence scores of the news topics for finding the optimal number of topics and then measure the similarity score between these two different languages news articles topics with the help of cosine similarity.

## Literature Survey

Now, discuss some fundamentals given by some of the other researchers in the direction of topic modeling by using LDA. Topic modeling is separate from rule-driven text mining methods using regular expressions or keyword matching techniques based on dictionaries. There are many techniques of topic modeling in which LDA is one of the most useful techniques.

**David M. Blei, Andrew Y. Ng et al** discussed about latent Dirichlet allocation (LDA). LDA is a generative probabilistic version for collections of discrete records along with text corpora and it is also a three-degree hierarchical Bayesian version [2]. In this research paper, every topic is modeled as an infinite combination over an underlying set of topic possibilities. The researchers have report outcomes in document modeling, text classification, and collaborative filtering, evaluating a mixture of unigrams model and the probabilistic LSI model.

**Hanna M. Wallach** was used text generation methods, which involve n-gram statistics, while others use latent topic variables assumed from the concept of "bag-of-words," where word order is avoided. In this research, the author discussed a hierarchical generative probabilistic model that combines both n-gram statistics and latent topic variables by expanding a unigram topic model to provide properties of a hierarchical bigram concept model in Dirichlet [7].

**Grant C. Atkins et al** define a methodology for determining the top news story for just a select group of U.S based online news websites and then measure correlations across them. To do that, they first created a headline and connect extractor that parses chosen blogs and then searched for ten US-based news site homepages for three months. Author's results show that they will configure synchronous articles besides related national events for a given day. This method could be used to further study the elections that are held from time to time [8].

**P. Fung** proposed a unique context heterogeneousness similarity live between words and their translations in serving to compile bilingual lexicon entries from a non-parallel English-Chinese corpus. The researcher had got shown initial results of matching words with their translations during an English-Chinese non-parallel corpus by victimization context heterogeneousness measures. Context heterogeneousness is often used as a clump live and discrimination lives. Its results are often used to bootstrap or refine a bilingual lexicon compilation algorithmic program [9].

**Vikas Thada and Dr. Vivek Jaglan** used cosine similarity, dice coefficient, Jaccard similarity algorithms. The analysis is based on the first 10 pages of the Google search

report and will be expanded to 30-35 pages for an effective performance estimate in future studies [10]. The cosine similarity was eventually assumed better performance for this dataset compared with others. Also, while the initial results are optimistic, there is still a long way to go to reach optimum crawling performance.

**M. Snover, Bonnie Dorr et al** explore a new way of using monolingual target data to improve the efficiency of a statistical or predictive machine translation for new stories. This technique employs comparable text—various texts in the target language which explore the same or equivalent stories as mentioned in the source language document. A broad monolingual dataset in the target language for each source document to be translated which is searched for documents that could be equivalent to the source documents [11]. The experimental results of this paper generated through the adaptation of the language and translation models show significant improvements over the baseline framework.

**Singh and Singh** compared three different approaches for estimating semantic similarity between two news items about (almost) the same topic/event to compare their similarity in two different languages. The experiment was tested using the Google News datasets [12].

## Data Preprocessing

Data collection for the same categories article in two different languages has always been a problem because of insufficient resources and the lack of publicly accessible corpus. The dataset used for this work is a news corpus that is collections of related categories news articles that are in two different languages (English & Hindi). After collecting the Hindi news, data firstly need to convert it in English by using "Google Translator". For this research, datasets were collected from "GoogleNews". When the dataset is available then start the preprocessing, which contains the tokenization tasks, remove all stopwords and process of creating Bigram. In this research, every keyword in the sentence is tokenized, and these tokenized keywords are then added to a Python list for the research purpose from the execution perspective. As we know, the English language has lots of stopwords. Those are also the words that are connected like prepositions and conjunctions. This research removes all stopwords by using NLTK (Natural Language Toolkit). After removing all stopwords, Bigram creation is an important part of the process of topic modeling. A bigram is a series of frequently occurring two adjacent tokens in the corpus [7]. Therefore, a probability for such keywords to appear one after another is calculated. When they have a sufficient value, then these pairs of words are merged and inserted in the dictionary as a new token. Bigrams are called as n-grams where n = 2. For bigrams, a conditional likelihood ($W_n$ given to $W_{n-1}$) is determined as follows:

$$P(W_n|W_{n-1}) = \frac{P(Wn, Wn-1)}{P(Wn-1)} \quad (1.1)$$

## Proposed Model

The goal of this work is to find ways to extract the topics from the given news corpus. An approach is introduced to finding out the correct topic to which news relates. This helps us to define news into its appropriate category.

The below Fig. 1 shows the fundamental layout of the workings of the model. When the dictionary gets available, apply the LDA algorithm with the preprocessing which is already done on it. This research has trained with 1000 news articles. The dictionary is also in the preprocessing period. Then, this entire dictionary runs into the LDA model and finds a variety of topics. Nevertheless, Latent Dirichlet Allocation doesn't understand how many topics are to find. So, to know the optimum number of topics, it needs to calculate a coherence score for each topic. And from this experimentation, we give the exact numbers of topics as a hyper-parameter in the trained model. The major issue in this modeling technique is that when too many topics are extracted it can get over fitted. So, to generate the correct number of topics is very necessary. Until we train our model and run the LDA algorithm, we run a coherence test with approximately 100 topics. Since we know that with about 1000 data, it is not possible to have 100 topics. So, we only fix the value to test the movement of incremental coherences through various topics then find that it reaches the highest around topic 40. Whenever this work has got the corpus in the model, we test it through experiments. This research also carried out a test of cosine similarity between two different languages (English and Hindi) news articles topics.

## Experimentation Result

The purpose of the first experimentation is to know the number of topics we have to derive from the topic modeling. The optimal number of topics cannot be recognized by LDA alone. We did an implementation to consider the optimal number of topics. This number depending on the dataset and the objective of the analysis. The goal of this research is to infer different languages (English and Hindi) of the same category online newspaper topics from about 1000 news instances. If so many topics are derived with an LDA model, it can overfit, which is not at all expected. But

on the other hand, extracting very few topics just do not make sense. So, a coherence dependent value for knowing the correct number of topics is considered (Fig. 1). This research worked with the model for both languages about 160 topics including the aggregated coherence value as seen in Fig. 2. In the below Figs. 2 and 3, along with the increment in the number of topics, firstly, the coherence score grows quickly and reaches its top when there are lists of topics near about 39 for English news articles and 40 for Hindi news articles. When this work has approximately 39 topics for English news articles and 40 for Hindi news articles topics, in this case, the research model works better according to the importance of its coherence score. Afterward, the coherence score drops gradually. This work executes for 160 topics for both languages because it is more than the range of topics that a newspaper is expected to have. At the top, this work has a coherence score near 0.61 for English news articles and 0.60 for Hindi news articles. So, this work was using 42 as its optimal number of topics for English news articles and 40 as for Hindi news articles.

## Similarity Measurement

Seeing as a learned LDA model also group topics in the form of many keywords. We are performing an experimentation to investigate cosine similarity measurements from the learned LDA model. Every time feed a few pairs of news to see the importance of cosine similarity. Although, Similarity calculation can also be achieved by using a Doc2Vec model. Ultimately, this work examines similarities between the LDA and the Doc2Vec. Table 1 indicates those scores.

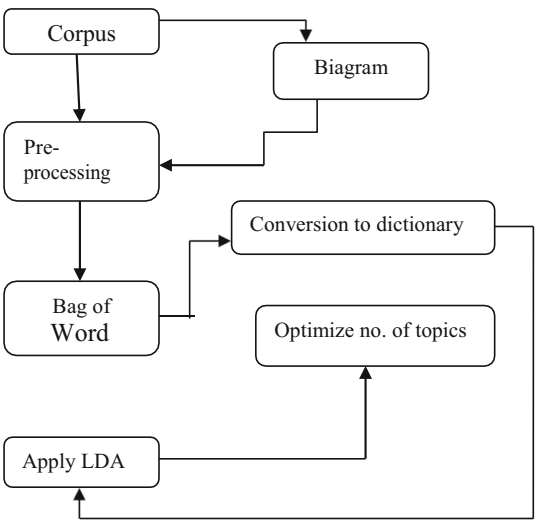


**Fig. 1** The proposed model of this research

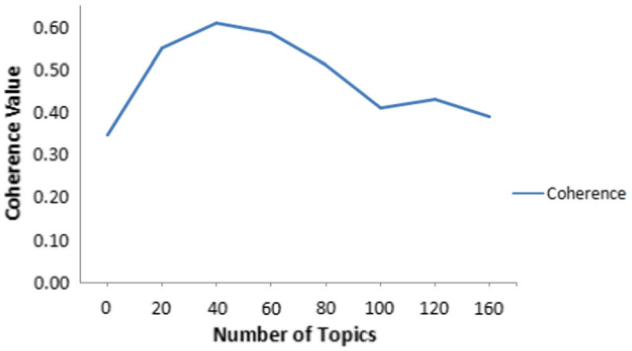


**Fig. 2** Coherence value for English news articles

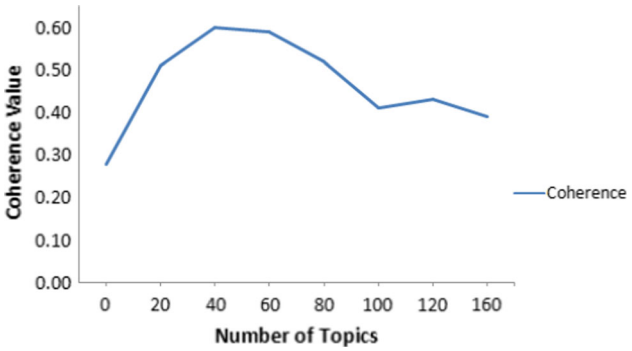


**Fig. 3** Coherence value for Hindi news articles

As this work has seen, a pair of two different languages (English & Hindi) news articles are loaded into these two models (LDA with Cosine & Doc2Vec) each time. For example, news article pair3 are very closely connected articles and talking about news "sport". In this way, news article pair1 are talking about the news "Corona Virus disease" which is a major problem in India. As a human interpretation, everyone can judge these two news pairs as a very closely connected pair. Cosine similarity with LDA provides these pairs in percentage news parir3 97.26% similar that should have been near to the human interpretation. Although Doc2Vec performed badly and only provided a similarity of 69.32%, this shows its bad outcomes. In this way, news article pair1 are similar to 93.21% with LDA cosine similarity, whereas Doc2Vec gives a similarity of 67.55%. This time also cosine similarity with LDA performs better as compare to Doc2Vec which is a winning scenario for LDA.

Also, this work has a measure of similarity with the HDP (Hierarchical Dirichlet Process) model that performs the worst of all of these. As we've seen in Table 2, HDP cannot catch in most cases any similarity at all. It does not relate the similarities among all the pairs apart from news pair3 are a strongly important pair in terms of its sentence and keyword frameworks. It didn't work out at all with all the other pairs. The System Mixture HDP isn't working.

**Table 1** Sample pair of same news article

| S.no. | Pair of English & Hindi news articles. |
|---|---|
| 1 | Corona virus \| 17 new corona virus cases reported in Jammu and Kashmir as of 5:00 PM - Apr 12<br>जम्मू-कश्मीर में तेजी से पांव पसार रहा कोरोना, मामले 17 और संक्रमित संख्या पहुंची 224 |
| 2 | Coronavirus \| British PM Boris Johnson discharged from hospital<br>अस्पताल से डिस्चाज हुए ब्रिटिश प्रधानमंत्री बोररस जॉनसन, कोरोना संक्रमण बाद कराया गया था एडमिट |
| 3 | Sachin Tendulkar interacts with 12,000 \| doctors on sport injuries<br>सचिन तेंदुल कर ने खेलों से जुडी चोटो पर 12,000 असपताल से साझा किए अनुभव |
| 4 | Vodafone - Idea launches cashback offer for online recharge done for other customers<br>वोडाफोन-आइडिया के तोहफा , हर रिचॉज पर मिलेंगे पैसे |
| 5 | Policeman's hand chopped off; two others injured in attack by 'Nihangis' in Ludhiana<br>पटियाल में निहांग सिखों का पुलिस पर हमला, एएसआई का हाथ काटा |

Let's investigate the normal estimation of these models to differentiate similarity and dissimilarity with the Fig. 4.

The factor LDA performs well because it can already identify the topics while the training time. Therefore, the topics help to identify the news article when this work scoring cosine similarity.

## Classifying News Article

This research developed an LDA topic model and derived topics for two different languages (English & Hindi) news, this work decided to go forward with the LDA news classification which are in two different languages. By using the LDA topic model, this work creates a technique for classifying news. Initially, this research obtains a document versus a topic matrix. Each word with a probability of belonging to a given topic is tagged in this matrix. Now, take an example for understanding this idea. Suppose, a document (D) = "Amitabh and Ashutosh are handsome." and ultimately become Dpreprocessed = ["Amitabh," "Ashutosh," and "Handsome"]. It can clearly understand as a human translator that this is a document of the topic "Person" and may also suggest making the topics k1 and k2. Moreover, the probability distribution matrix for document v/s term will look as the below matrix:

**Table 2** Showing comparison study for cosine similarity score between two different models

| News article pair (Docs | Cosine similarity with LDA | Doc2Vec | HDP |
|---|---|---|---|
| 1 | 0.9321 | 0.6755 | 0.1452 |
| 2 | 0.8215 | 0.3672 | 0.0221 |
| 3 | 0.9726 | 0.6932 | 0.2747 |
| 4 | 0.8021 | 0.4172 | 0.0111 |
| 5 | 0.7925 | 0.3823 | 0.0000 |

**Table 3** Matrix table

| | | |
|---|---|---|
| (W1, P1) | (W1, P2) | (W1, P3) |
| (W2, P4) | (W2, P5) | (W2, P6) |
| …. | …… | ……. |
| …… | …… | …… |
| …… | …… | …… |
| ……. | …… | …. |
| (Wn, Pk-2) | (Wn, Pk-1) | (Wn, Pk) |

| | |
|---|---|
| (0, P1) | (0, P2) |
| (1, P3) | (1, P4) |
| (2, P5) | (2, P6) |

In this above matrix, P1, P2, P3, P4, P5, P6 are probabilities, and 1, 2, 3 are indexes of words from the above example.

Therefore, we get;

$$\Sigma(P1 + P2) = \Sigma(P3 + P4) = \Sigma(P5 + P6) = 1 \quad (1.2)$$

In the research objective, for each topic, this work calculates the mean of the keywords. Now, suppose there are n-terms and k-topics so model looks like as (Table 3).

Therefore, the mean (k) topic for each word is determined as the following equation.

$$\text{mean1} = \Sigma(P1.....P3)/k \quad (1.3)$$

$$\text{mean2} = \Sigma(P4.....P6)/k \quad (1.4)$$

$$\text{mean}(k) = \Sigma(Pk - 2.....Pk)/k$$

Thus, the news articles related to the topic with the highest value of probability.

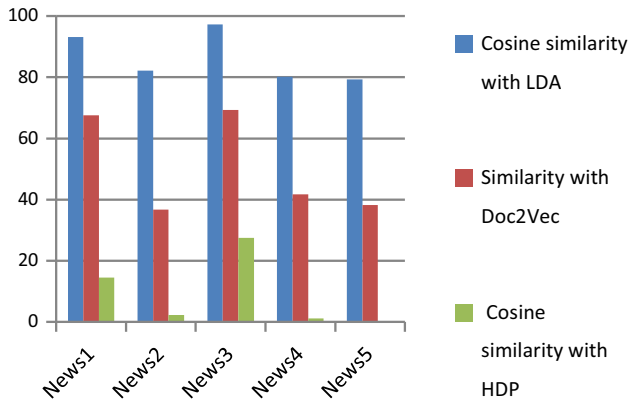


**Fig. 4** Comparison of similarity coefficients for articles of the same news

## Extracting Topics

As follows, Figures 5 and 6 describe those topics extracted from the English and Hindi newspaper of the same category. As we can see, in both languages news documents, every topic contains 10 keywords relating to the topic and that keywords have specific probabilities. Also, this work shows 10 greatest probability keywords for every topic. Moreover, few keywords may have no significance for this topic but many of them are important. Many topics are very confused and the definition could not be the same, except many of the other topics cause senses. Although, The topics aren't automatically tagged. The tags are created manually by putting these keywords groups for every topic, as LDA can only include the category of words defined as topics. This research has got probabilities in a decreasing order with each term. This work was using a news article about a Borish Johnson's health as shown in Figs. 5 and 6.

We used a text about Borish Johnson in Figs. 5 and 6 this is how the topic appears like distribution. Topic 16 is the most important topic for the Hindi news article and Topic 23 is the most important topic for the English news article which outperforms the other topics in aspects of their score of probability.

Although the news is about Borish Johnson's health for both languages, the model is giving us such a valid response. It provides us a topic that is about Borish

```
16,
  '0.029*"Borish_Johnson"+ 0.026*"save" +
0.027*"life"+0.024*"persons"+0.020*"Satur
day" + 0.019*"post" + 0.017*"discharged"
+0.012*"coronavirus"+0.011*"treatment"+0.
011*"intensive_care_unit"'),
```

**Fig. 5** Word clusters for English news topic: Borish Johnson's health

```
(23,
  '0.028*"life" + 0.026*"save" +
0.025*"Johnson" + 0.024*"people" +
0.019*"post" + 0.017*"medicament" +
0.017*"discharge" + 0.012*"COVID-19" +
0.013*"day" + 0.012*"ICU"')
```

**Fig. 6** Word clusters for Hindi news topic: Borish Johnson's health

Johnson's health-related words. This experiment reflects the fact that this work has got the core topic from both languages news successfully with this model.

## Conclusion

Topic models have seen several achievements in past years and it is used in several applications like document tag assignment, topic-based search applications, and also for document summarization. In this work, we have shown how topic modeling can be applied to different languages (English Hindi). The ongoing research analyzed to finding the core topics of the same category news article which is in two different languages (English and Hindi) and then also measures the similarity score between these different languages topics. In the proposed work, this research used "GoolgeNews'' for collecting the datasets. The three approaches are cosine similarity with LDA, HDP, and Doc2Vec. In these three approaches, cosine similarity with LDA performed better than Doc2Vec and HDP and gives the highest similarity score when this work is finding the similarity between those news article topics. For future work, we may use polyLDA and LSI (a modification of LDA) in the research study, and specific similarity measures for classifying other languages may also be discussed.

**Funding** No funding was received to perform this research.

**Declarations**

**Conflict of interest** The Authors have no conflicts of interest related to the content of this study.

## References

1. "How Much Data Does The World Generate Every Minute?" [Online]. https://www.domo.com/news/press/how-much-data-does-the-world-generate-every-minute
2. D.M. Blei, Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)

3. M. David Mimno, M.W. Hanna, N. Naradowsky, A.S. David, "Polingual Topic Models," vol. Proceedings of the 2009 Conference on Empirical Methods, pp. 880–889, August (2009).
4. MB. David, YN. Andrew, IJ. Michael, Latent Dirichlet Allocation. J. Machine Learning Res. **3**, 993–1022 (2003)
5. R. Alghamdi, K. Alfalqi, A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **6**(1), 147–153 (2015)
6. H.Z.Z. Tong, A text mining research based on lda topic modelling. Jodrey School Comput Wolfville NS **10**, 201–210 (2016)
7. M. Hanna "Topic Modeling beyond bag-of-word," pp. 977–984, (2006).
8. G. Atkins, M. Weigle, M. Nelso, "Measuring News Similarity Across Ten U.S. News Sites," pp. 1- 11, June (2018).
9. P. Fung, "A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora," pp. 173–183, 1529 (1995).
10. VJV. Thada, "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm," *International Journal of Innovations in Engineering and Technology (IJIET)*, vol. 2, pp. 202–205, August (2013).
11. B. Dorr, RSM. Snover, "Language and Translation Model Adaptation using Comparable Corpora.," *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.*, pp. 857–866, October (2008).
12. S. Singh, R. Singh. Text Similarity Measures in News Articles by Vector Space Model, "*The Institution of Engineers* (*India)*, pp.329–338 (2020).