Vol. 04, No. 2,(2020),078–091, **ISSN:2581-3242**

IJMLNCE JOURNAL

International Journal of Machine Learning and Networked Collaborative Engineering

Journal Homepage: http://www.mlnce.net/home/index.html

DOI : https://doi.org/10.30991/IJMLNCE.2020v04i02.004

Discovering Trending Topics from the Tweets by Odia News Media during Covid-19

[a]Swarupananda Bissoyi[*], [b]Brojo Kishore Mishra, [b]Raghvendra Kumar

[a]Dept. of Computer Application, North Orissa University, Baripada, Odisha, India
[b]Department of CSE, GIET University, Gunupur, Odisha, India

[a]swarupananda.bissoyi@nou.nic.in, https://orcid.org/0000-0002-3995-6857

[b]bkmishra@giet.edu, https://orcid.org/0000-0002-7836-052X

[b]raghvendra@giet.edu, https://orcid.org/0000-0003-1311-7585

Abstract

The onset of the Covid-19 pandemic and the imposed lockdown has fueled the news consumption significantly. News portals, including the ones in Odia language, actively feed news related to Covid-19 to their consumers via their websites and Twitter handles. The news items didn't restrict to Covid-19 alone; they also touched various life domains like education, healthcare, administration, politics, movies, etc. Discovery of the news trends provides a holistic view of the issues and topics popular in the online community. This could be of interest to advertisers, marketers, researchers, sociologists, and policymakers. This paper applies Topic Modeling to discover the trends from the tweets made by the Odia news media from 20th March 2020 to 31st August 2020, the period which saw the emergence of both 'lockdown' and 'unlocks' in India. We found that during this period, the Odia news media didn't restrict themselves to report news surrounding Covid-19; instead, they said other happenings as well.

Keywords

Trend Analysis

Topic Modeling

Twitter

Odia News Media

Covid-19

1. Introduction

In recent years, Twitter has established itself as one of the major social networks and microblogging services. The Covid-19 pandemic onset has resulted in an infodemic in Social media, and Twitter is no exception. Many news agencies and portals have used Twitter as a potential medium to disseminate information to their consumers. Twitter has gained researchers' attention due to the copious amount of data generated every second based on a broad number of topics. Researchers have been long focusing on the analysis of tweets to get meaningful insights and present trends. Discovering the trending issues, real-world event detection, sentiment analysis, automatic summarization, Spam detection, Social behavior analysis, etc. are some of the critical research areas concerning the data generated by Twitter. These research focuses are not limited only to the English language but also in many popular languages worldwide. Twitter is conceived both as a news media and a social network [1]. Most of the famous news media houses today break their news first on Twitter. Of late, the news media in Odia language have started this practice and have begun tweeting in Odia language to reach out to their potential readers.

The Odia language is the 6th Indian language and the first amongst the Indo-Aryan language to get 'Classical Language' status by the Indian government due to its long history in literature and originality. Despite being a 1500-year-old language, the Odia language is yet to find a strong foothold in the Internet and

Digital Media. It's only after the Windows 7 operating system where Odia Unicode was fully supported; users could practically use Odia language on the web for their day to day communication. Further, in recent years, support of the Odia language in almost all the mobile phones and cheaper data rates has fueled the growth of the use of the Odia language over the web, news portals, social media like Twitter and Facebook. Today, as many as 58 news portals are in Odia language, 38 out of them regularly push their news items on Twitter. News trend analysis deals with news topic discovery, predicting news trends, and understanding social behavior patterns. This has gained significant traction due to its usefulness in contextual advertising, building recommender systems, and document classification tasks. Twitter itself has a "What's happening' section in its portal where the trending topics are shown based on user's location, user followings, personal profile, user's domain interests like politics, sports, etc. Twitter's algorithms discover these trending topics based on the tweets' volume and velocity concerning a subject, hashtag analysis, coverage, popularity, reputation, and tweeting behavior of the users. Other researchers also have delved into this news trend analysis by suitably using Twitter API to fetch Twitter data and performing analysis on top of it. So far, the trend analysis is being done with traditional NLP techniques like n-Gram based relative normalized TF-IDF analysis, classification by machine learning, and frequent pattern mining. LDA-based Topic modeling [2] is being applied to discover the tweets' trending topics automatically. Also, many researchers have combined it with temporal information to identify the evolution of issues over time [3], [4], [5], [6].

Tweets can be viewed as a collection of documents, and Topic Modeling can be used to derive latent topics by processing the tweets as documents. Topic Modeling is a statistical method, mostly used to extract representative latent cases from a corpus. Topic models are probabilistic models representing topics as multinomial distributions over words, with a premise that each document from the corpus is described as a suitable mixture of topics. To the best of our knowledge, Topic Modeling is never tried before on Odia language texts. This paper's contribution is about discovering the trending topics by applying Topic Modeling out of the tweets made in Odia language by the news agencies during the Covid-19 period. The period chosen is from 20th March 2020 to 31st August 2020, the period which saw the emergence of both 'lockdown' and 'unlocks' in India. The finding's crux is that the Odia news media didn't limit themselves to Covid-19 news and considerably reported news other than Covid-19.

The paper is constituted as follows: First, we study the related works in this area regarding Trend Analysis and Topic modeling on social media data, particularly Twitter. We then discuss methods of collecting tweets from Odia news media, preprocessing Odia texts, and three different Topic Modeling algorithms applied to them. We evaluated the algorithms in terms of the Topic Coherence score. Finally, we showcased our results and visualized the topic models to identify trending topics.

2. Background

Topic modeling is a method of unsupervised classification used to extract unrevealed topics in a textual database. It is a statistical model used for text mining. It identifies the hidden pattern in a corpus of unstructured textual data. The concept of Topic Modeling was first introduced using Latent Semantic Indexing (LSI). This non-probabilistic model constructs a semantic space by identifying latent relationships within a collection of documents [7]. LSI is later popularized by Papadimitriou et al. [8]. Thomas Hoffman contributed a probabilistic model for Topic Modeling, i.e., PLSA (Probabilistic latent semantic analysis) [9]. David M. Blei et al. introduced another probabilistic model called "Latent Dirichlet Allocation" (LDA), an algorithm in the realm of machine learning [2], which became very popular and it's the most commonly used topic modeling algorithm today. Dynamic Topic Model is a generative model augmentation to the LDA proposed by Blei et al. that can determine how undiscovered topics are unfolding from a corpus over time [3]. Hierarchical Dirichlet Process (HDP) is again an LDA augmentation, which deals with the number of topics unknown beforehand. A Dirichlet process is used to determine the appropriate count of the mixture components [10].

Many researchers have applied Topic Modeling on Twitter data. Surian et al. worked on characterizing the discussions made on Twitter about HPV vaccines, using topic modeling. They observed two years of Twitter posts on HPV vaccines using LDA and DMM (Dirichlet Multinomial Mixture) to determine the topics associated with the tweets [11]. Ghosh and Guha used LDA and GIS spatial analysis to identify the tweets people making on obesity and the common obesity-related themes [12]. Melis and Savesk proposed a pooling technique and LDA and Author Topic Model (ATM) to aggregate the tweets in the same user-to-user conversation [13].

In the analysis of news trends from the tweets, Lu and Yang augmented the Moving Average Convergence-Divergence (MACD) indicator with a trend momentum parameter to predict news trends in Twitter [14]. Mathioudakis and Koudas developed TwitterMonitor to perform real-time discovery of trending topics from the tweets. They identified high occurring keywords and grouped them in terms of co-occurrences [15]. Lau et al. proposed a topic modeling approach to detect trending events on Twitter. They processed the tweets online while updating the topic model based on time slices, thereby dynamically changing the vocabulary [16]. Many interesting research pieces have come up out of the analysis of tweets in the Covid-19 pandemic period. Sha et al. used the 'Hawkes Binomial Topic Model' on US lawmakers' tweets to identify the subtopics about risk, testing, and treatment of Covid-19 [17]. De Santis et al. analyzed a system for identifying pertinent topics out of the posts made by Italian Twitter users by employing a pipeline of NLP, Content Aging Theory framework for determination of qualitative parameters of the tweets, co-occurrence analysis to build a topic graph, and graph partitioning [18]. Singh et al. analyzed the conversations made on Twitter worldwide and suggested the existence of a Spatio-temporal relationship between flow information and the new cases of Covid-19. They further discovered that myths and misinformation surrounding Covid-19 are insignificant compared to the other conversations [19]. Lwin et al. performed a trend analysis of four sentiments viz. 'fear,' 'anger,' 'sadness,' and 'joy' by analyzing over 20 million tweets made during Covid-19. They pointed out the rise and fall of these sentiments in the user community over the different governments' different phases dealing with Covid-19 [20]. Ordun et al. applied pattern matching and topic modeling on tweets related to Covid-19 to identify 20 distinct topics. They further applied Uniform Manifold Approximation and Projection to determine the quality of topics. They also analyzed retweets to understand the propagation of information [21]. Kabir and Madria created a dataset of tweets generated from US and applied Topic modeling using LDA to perform a temporal study of topic-changes, subjectivity, and emotions [22].

There is a significant paucity of research on NLP in the Odia language when it comes to the Odia language. Though many researchers have tried basic NLP tasks on Odia language texts like Stemmer Development [23], [24], Spell Checker [25], Morphological Analyzer [26], [27], [28], and Named-Entity Recognizer [29], [30], they have not resulted in any tools for future research in the public domain. In the domain of news trend analysis, we could only identify three research publications. Jena and Mohanty categorized the news articles as' positive,' 'negative,' and 'neutral' using 'Syntactosemantic' tagging. They further performed a sensitivity analysis of Odia news articles by training on an SVM classifier of tf-idf vectors obtained from the text's unigram and bigrams [31]. The same authors in a similar work predicted the public opinion from the Odia newspaper articles [32]. Mohanty et al. by annotating a corpus of Odia sentences collected from the newspaper "The Samaja" and performed Sentiment Analysis using the 'SentiWordNet' developed by them [33]. To the best of our investigation and exhaustive survey of the literature, we find that neither Topic Modeling is tried on an Odia language corpus nor any analysis of the tweets in Odia language was performed. This speaks about the novelty of this work.

3. Methodology

In this work, we will be using LDA governed Topic Modeling to discover the trending topics from the tweets. Our methodology consists of five distinct stages, as shown in Figure 1, which are Data collection, Preprocessing, Topic Modeling using LDA along with LSI and HDP, Evaluation between LDA, LSI, and HDP, and finally, Visualization of the Topic model for proper interpretation. A detailed description of the stages follows in the next section.

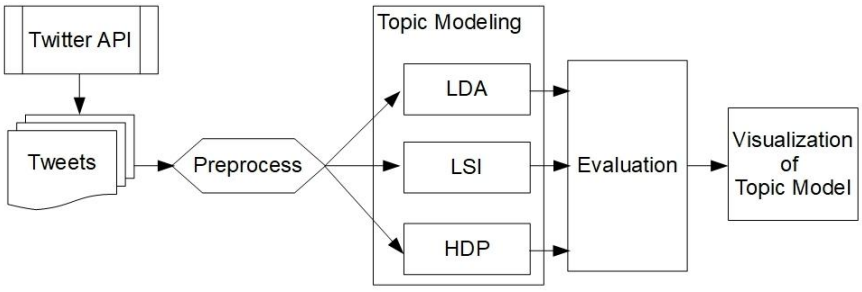

Figure 1. Process Employed for Topic Modeling on tweets

3.1. Data Collection

We have used the REST API by Twitter to get the tweets following the 'Application only Authentication' method, making API requests on its behalf, without the user context. We collected the tweets made between 20th March 2020 to 31st August 2020 by 30 popular Odia news Twitter handles viz. SamajaLive, sambad_odisha, NEWS7Odia, DharitriLive1, nitidintoday, otvnews, kanak_news, News18Odia, NandighoshaTV, theargus_in, PratinidhiOdia, ZeeOdisha, sancharlive, odishabhaskar, OKhabara, Odisha_Sambad, odisha_kranti, knewsodia, OdishaReporter, OdishaLink, SatyaPrNayak, odishasamachar, OdiaSpot, onakhabar, sakalakhabar, odishatime, aetajeevan, fastnewsnetwor1, OdiaScraps, and OdishaLive. We got a total of 27,318 tweets made by those Twitter handles.

Table 1. Tweeting patterns of the Odia news media

URL Links	Hashtags	Mentions	Twitter Handles
Yes	No	No	OdishaLive, fastnewsnetwor1, odishatime, OdiaSpot, odishasamachar, nitidintoday
Yes	Yes	No	OdiaScraps, OdishaLink, knewsodia, odisha_kranti, odishabhaskar, NEWS7Odia, SamajaLive
Yes	No	Yes	aetajeevan, onakhabar
Yes	Yes	Yes	SatyaPrNayak, OdishaReporter, Odisha_Sambad, sancharlive, ZeeOdisha, PratinidhiOdia, theargus_in, NandighoshaTV, News18Odia, kanak_news, DharitriLive1, sambad_odisha

Almost all the Twitter handles tweeted the headlines they reported in their corresponding news portal website. The pattern is like 'Headline' followed by the shorted URL link. The design is almost universal across all those news media. However, few of them did use hashtags and mentions (apart from self). Table 1 shows the patterns, and corresponding Twitter handles.

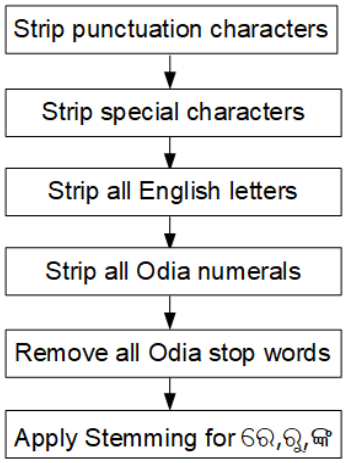

Figure 2. Steps of Preprocessing

3.2. Preprocessing

Preprocessing is necessary for making input data amenable for subsequent analysis. Since Tweets are noisy sentences with various punctuation marks, special characters used to shorten the texts, hashtags, etc., preprocessing is necessary to normalize the entries in the dataset. The preprocessing steps we followed are depicted in Figure 2. We removed the punctuation characters and special characters. We then removed all the English letters, digits, and all Odia numerals. We identified around 370 stop words in the Odia language,

which we removed from the text. Finally, we performed a basic stemming on the words, with the removal of suffixes like 're'(ರೆ), 'ru'(ರು), and 'nka'(ಗಳ).

3.3. Topic Extraction

For topic extraction, we used the Latent Dirichlet Allocation (LDA). With Bayesian inference as to its foundation, LDA is a three-layered hierarchical model in which each of the constituent documents of a document collection can be viewed as a finite blend over a group of hidden topics. LDA visualizes every document as a collection of different topics having a certain probability. The sparse 'Dirichlet priors' in the allocation signifies that the documents span a limited set of topics only and each of the topics is a limited set of frequent words. It brings about a superior disambiguation of words and a better correlation of documents to the topics than PLSA, which is based on mixture decomposition.

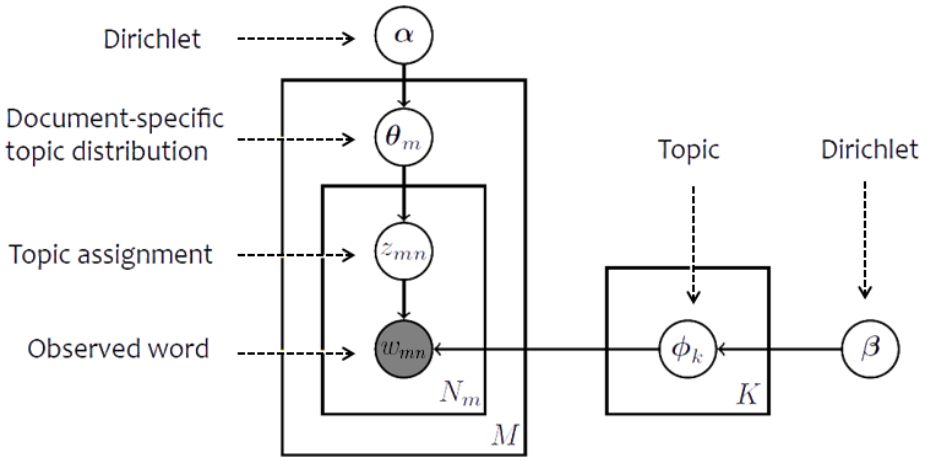

Figure 3. LDA plate notation showing Dirichlet distribution of topic-word distributions

Figure 3 shows the LDA Topic Model using the plate notation [34]. In plate notation, popularly used in Bayesian inference, a plate or rectangle is used to represent variables that repeat in a sub-graph and the number in the plate represents the number of repetitions. In this figure, the outer plate corresponds to the documents, whereas the inner plate corresponds to the repeated word positions within a given document. The figure shows the Dirichlet distribution of topic-word distributions of M documents, N words in a document, and K topics. α is the 'Dirichlet prior' parameter for the topic distribution per document. Higher the value of alpha would mean each document is more likely to contain a blend of most topics instead of any single specific topic. β is the 'Dirichlet prior' parameter for the word distribution per topic. Higher the value of beta would mean each topic is more likely to contain a blend of most words instead of any specific word. θ is the distribution of topic for the document m, ϕ is the distribution of words for the topic k, x_{mn} is the specific word and z_{mn} is the topic related to the *n*th word in the document *m*. Here only W is the observable variable; except W all others are latent variables. The Dirichlet-distributed topic-word distributions are denoted by $\phi_1,\ldots,\phi_K$.

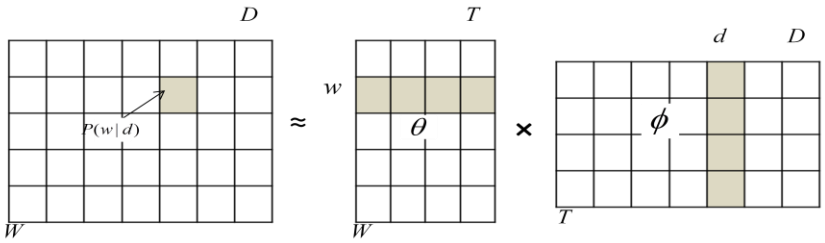

Figure 4. Matrix of words in a document composed of two matrices consisting of distribution of words in a topic and distribution of topics in a document

The original document-word matrix can be viewed as a composition of two matrices θ and ϕ. The matrix θ is the matrix of documents (as rows) and topics (as columns). The matrix ϕ is the matrix of topics (as rows) and words (as columns). Therefore, $\theta_1, \ldots, \theta_M$ is the set of vectors each representing topic-distribution and $\phi_1, \ldots, \phi_K$ is the set of vectors each representing word distribution. Therefore,

$\theta = P(t \mid d)$ i.e., distribution of 'topics' in documents

$\phi = P(w \mid t)$ i.e., distribution of 'words' in topics

and the probability that a word w belongs to a document d is given by:

$$P(w \mid d) = \sum_{t \in T} P(w \mid t) P(t \mid d)$$

which is the dot product of θ_{td} and ϕ_{wt} for each topic t. This is shown in matrix form as shown in Figure 4.

Algorithm 1 Generative Process

Generate(K, D, M, N)
{
  foreach topic $k \in 1, ..., K$ {
    $\phi_k \sim Dirichlet(\beta)$ // word distribution per topic.
    foreach document $m \in 1, ..., M$ {
      $\theta_m \sim Dirichlet(\alpha)$ // distribution over topics
      foreach word $n \in 1, ..., N_m$ {
        //draw topic assignment
        $z_{mn} \sim Multinomial(1, \theta_m)$
        $w_{mn} \sim \phi_{z_{mn}}$ //draw word
      }
    }
  }
}

To find the latent topics in a set of unknown documents, a generative process is employed wherein, given a set of topics how possibly the documents could have been created. Once we know the process, maybe we can reverse engineer to discover the topics given a set of random documents because documents are nothing but the random blend of latent topics, wherein each of the topics is potentially a distribution over all the words. The generative process for a set of document corpus D consisting of M documents each of length N_i is shown in Algorithm 1. The generative process results in the joint distribution as follows:

$$P(w, z, \theta, \phi \mid \alpha, \beta) = P(\phi \mid \beta) P(\theta \mid \alpha) P(z \mid \theta) P(w \mid \phi, z)$$

The ultimate aim of LDA is estimating θ and ϕ. Estimating θ is about finding which words are significant for which topic and estimating ϕ is about finding which topics are significant for a given document. What we are interested in, is the unobserved latent parameters z, θ and ϕ. Each θ_d is a document representation in topic-space, and each z_i corresponds to the topic which generated the word w_i, and $\phi_{i,j} = P(w_i \mid z_j)$.

Algorithm 2: Topic Modeling algorithm

```
LDA(tweets, k)
{
  k ← Number of topics to be categorized
  foreach tweet in the tweets {
    construct bagOfWords from the tweet
    foreach word in the bagOfWords {
      t ← randomly chosen topic out of k topics
      assign a topic t to word
    }
  }
  /*Now all the tweets are distributed with k topics and all
    topics are distributed over all the words but that may not be
    the correct representation. Therefore we improve it as
    follows:*/

  foreach tweet d in the tweets {
    foreach word w in d {
       compute p(t|d)  /* portion of words in the tweet d that
                          gets assigned to topic t */
       compute p(w|t)  /* portion of the assignments to topic t,
                          over the tweet d, that originated from
                          the word w */
       Reassign w, to a newly formed topic t', where the topic
       t' is chosen having a probability p(t'|d)*p(w|t')
    /* Reiterating the last step a significant number of times
    would result in a consistent state where the assignment of
    topics is steady and generate an overall decent mixture of
    topics in each of the tweets. */
    }
  }
}
```

3.4. Learning the Posterior Distribution

The key challenge in Topic Modeling is posterior inferencing i.e., learning posterior distributions of the latent variables given the observed data. This is nothing but the generative process in reverse. In LDA, this is done by solving the following:

$$P(\theta, \phi, z \mid w, \alpha, \beta) = \frac{P(\theta, \phi, z, w \mid \alpha, \beta)}{P(w \mid \alpha, \beta)}$$

The issue here is that there is no efficient algorithm to compute the above distribution. The denominator $P(w \mid \alpha, \beta)$ which is the normalization factor cannot be computed precisely. However, many approximate inference methods can be applied, one of which is Collapsed Gibbs Sampling (a variation of Gibbs Sampling), which we will be using for our LDA Topic Modeling [35].

3.5. Topic Extraction from the tweets

The final algorithm used for Topic Modeling on tweets is listed in Algorithm 2. The algorithm uses a bag of word models. The assignments of the topics follow the Collapsed Gibbs Sampling algorithm. We applied the Topic Modeling algorithm to the tweets collected and preprocessed. We set the number of topics to be extracted as six. The results and Visualization are shown in the next section.

Table 2. Six distinctly interpretable topics from the tweets along with the percentage of total tokens in the dataset that the topic relates to.

Topic1 (12%)	**Topic2** (23%)	**Topic3** (14%)	**Topic4** (15%)	**Topic5** (19%)	**Topic6** (17%)
ସହିଦ (sahid)	କରୋନା (karonā)	ପରୀକ୍ଷା (parīkṣā)	ଦିବସ (dibasa)	ସୁଶାନ୍ତ (suśānta)	ଜିଲ୍ଲା (jillā)
ମୁକାବିଲା (mukābilā)	ଆକ୍ରାନ୍ତ (ākrānta)	ଭାରତ (bhārata)	ରଥ (ratha)	ମାମଲା (māmalā)	ବର୍ଷା (barṣā)
ସ୍ଥିତି (sthiti)	ଚିହ୍ନଟ (cihnaṭa)	ପ୍ରଧାନମନ୍ତ୍ରୀ (pradhānamantrī)	ଜିଲ୍ଲାକୁ (jillāku)	ମୃତ୍ୟୁ (mṛtyu)	ଗାଁ (gāṁ)
ରାଜଧାନୀ (rājadhānī)	ରାଜ୍ୟ (rājya)	ମୋଦି (modi)	ଓଡ଼ିଶାର (oḍiśāra)	ମୁଖ୍ୟମନ୍ତ୍ରୀ (mukhyamaṃtrī)	ବନ୍ୟା (banyā)
ମୁଣ୍ଡ (muṇḍa)	ପଜିଟିଭ (pajiṭiv)	ଦେଶ (deśa)	ଅଭିନେତା (abhinetā)	ସରକାର (sarakāra)	ପ୍ରବଳ (prabala)
ଯବାନ (yabān)	ସୁସ୍ଥ (sustha)	ସିଂ (sing)	ମୃତ (mṛta)	ଟଙ୍କା (ṭaṅkā)	ଢେଙ୍କାନାଳ (ḍheṅkānāla)
ମଦ (mada)	ହଜାର (hajāra)	ବିପଦ (bipada)	ମା (mā)	ମନ୍ତ୍ରୀ (mantrī)	ଉଦ୍ଧାର (uddhāra)
ସହର (sahara)	ନୂଆ (nūā)	ନୂଆ (nūā)	ଗୁରୁତର (gurutara)	ବନ୍ଦ (banda)	ପାଣି (pāṇi)
ମନା (manā)	ମୃତ୍ୟୁ (mṛtyu)	ଆମେରିକା (āmerikā)	ପୁଅ (pua)	ସୁପ୍ରିମକୋର୍ଟ (suprīmkorṭ)	ଆଜିଠୁ (ājiṭhu)
ଘୋଷଣା (ghoṣaṇā)	ବୃଦ୍ଧି (bṛddhi)	ଭାରତୀୟ (bhāratīya)	ଦୁର୍ଘଟଣା (durghaṭaṇā)	ଶେଷ (śeṣa)	ଫିଲ୍ମ (philm)

4. Results

Table 2 shows six distinctly interpretable trending topics extracted from the tweets with the 10 individual topic terms sorted by the weights in descending order. The table also shows the percentage of total tokens in the dataset that the topics correspond to. Topic2, which is about Corona, identification of positive cases, recovery, etc. is the most trending topic, with 23% of the total tokens. Figure 5 shows the importance of the representative words per topic. Each of the sub figures corresponds to a topic with the representative words plotted on the x-axis and the word count and the weight of the word for that topic are shown in the y-axes. The plot corresponding to the topic identified by 'Topic2' is having the word Corona ('କରୋନା') as the most frequent word with a word count of 6,049 and a term weight of 0.131147 for that topic. However, the topic distribution shows that, apart from Covid-19, other burning topics of that time did surface with the news of the death of Sushant Singh Rajput resulting in Topic5, the news about martyred soldiers in the clash with Chinese troops at LAC resulting in Topic1, the news pertaining to flood situation of Odisha resulting in Topic6, the national news coverage regarding India covered in Topic3, and the local news reporting various road accidents occurred during that time covered in Topic4.

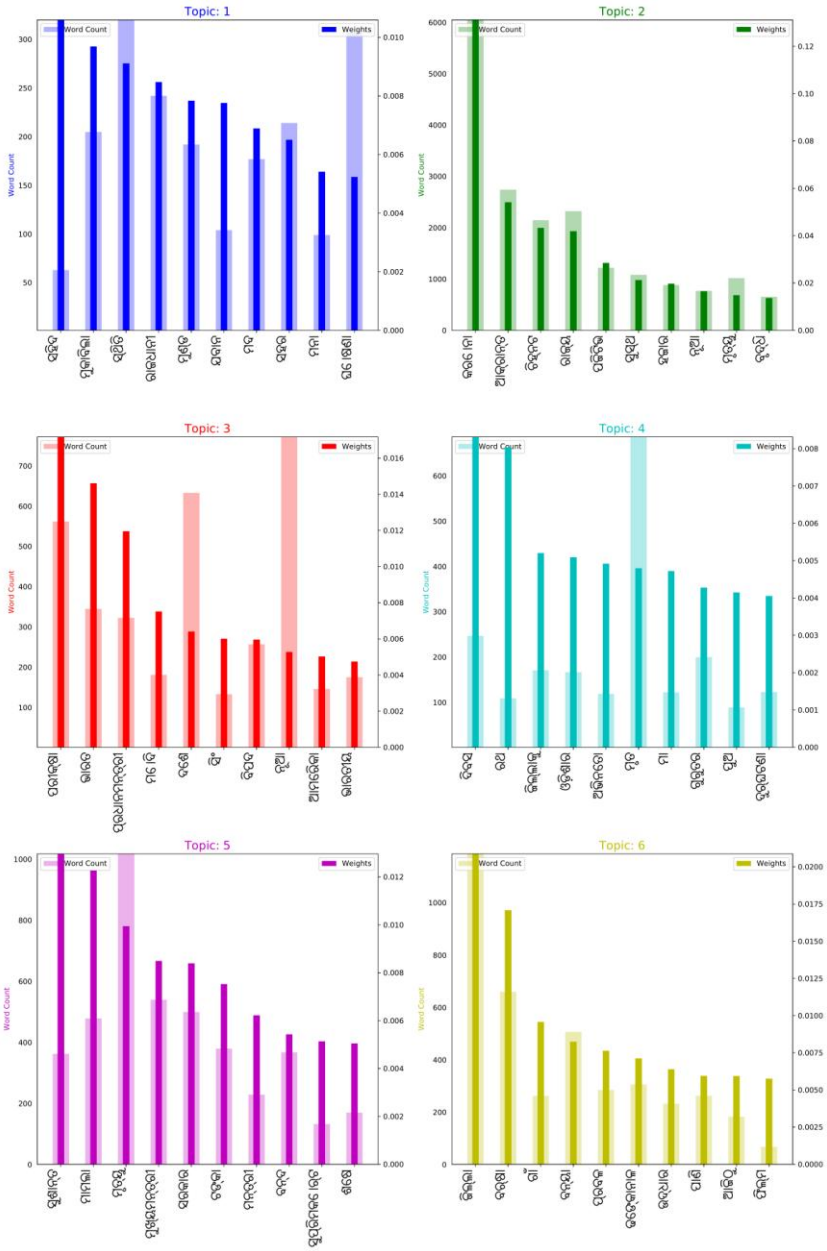

Figure 5. Topic Representation in terms of Word Count and Topic Importance

4.1. Visualization

We used pyLDAvis, a Python library intended for interactive Visualization of the generated topic model. Figure 6 shows the Top 30 salient terms. They are shown in descending order in terms of their 'Saliency' [36]. The Saliency of a term helps in determining whether that term is useful in distinguishing the topic. It's computed as:

$$\text{saliency}(w) = P(w) \cdot \sum_{T} (P(T \mid w) \cdot \log \frac{P(T \mid w)}{P(T)})$$

where $P(T \mid w)$ is the probability of a topic given a word w, and $P(T)$ is the topic distribution over the entire corpus.

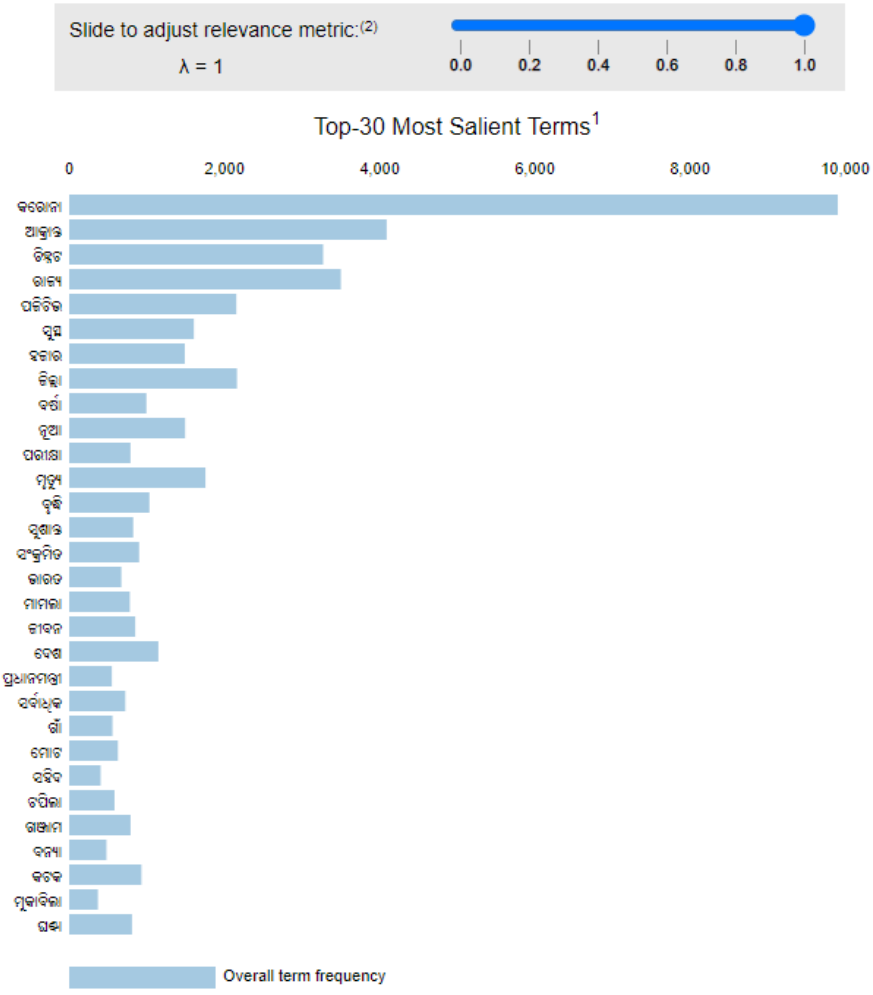

Figure 6. Top 30 salient terms

Among the 30 words shown, the word 'Corona' (କରୋନା), found in Topic2 is the most salient word with the term frequency touching more than 10,000. It's evident that the news trends of Odia news media during the said period of Covid-19 were mainly around news about Corona, the death of Sushant Singh Rajput, and the flood situation in Odisha.

5. Evaluation

We evaluated our model using the Coherence score [37], a measure used for determining the quality of the topic learning. The higher the Coherence better is the Topic model. Topic coherence is calculated as the sum of pair-wise distributional similarity scores over the topic words, W i.e.:

$$Coherence(W) = \sum_{(w_i, w_j) \in W} Score(w_i, w_j, \sigma)$$

where σ is a smoothing factor. The score is computed as the point-wise mutual information (PMI) for a pair of words, i.e.,

$$Score(w_i, w_j, \sigma) = \log \frac{p(w_i, w_j) + \sigma}{p(w_i)\, p(w_j)}$$

where $p(w_i)$ is the probability of w_i in a random document, and $p(w_i, w_j)$ is the probability of co-occurrence of both w_i and w_j in a random document.

We computed the Coherence score for the LSI model, HDP model, and LDA model. The Coherence scores are plotted in Figure 7. LDA model's coherence score is computed as 0.42876, which is higher than that of the HDP and LSI model. Therefore, it's evident that LDA is better than both HDP and LSI model.

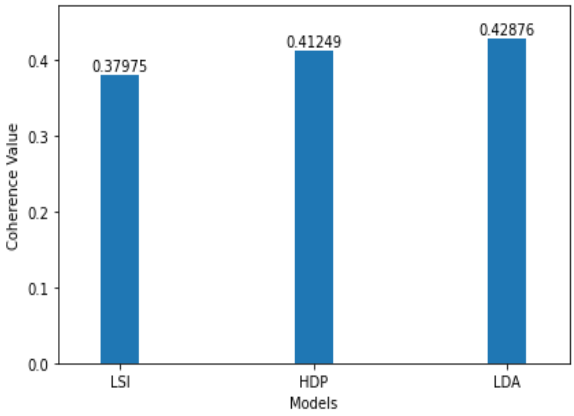

Figure 7. Coherence values were computed for LSI, HDP, and LDA models.

6. Conclusion

In this paper, we have successfully extracted the trending topics out of the tweets made by Odia news agencies during the Covid-19 period. The distinctiveness of the topics shows the effectiveness of the topic model. We have applied basic NLP techniques to preprocess the tweets to make it amenable for topic modeling. We found the trending topics discovered from a little less than 30,000 tweets made in Odia language, in line with perception built-in public as news trends. Moreover, Odia news media tweets during the Covid-19 pandemic period didn't restrict the news surrounding Covid-19, but they did cover other news of importance as well. The model's quality can be improved with a larger dataset and better NLP tools for the Odia language. In the future, we will be applying it on a larger dataset of tweets in Odia language along with other significant NLP tasks for Odia language, such as the development of an electronic lexicon, Stemmer and Lemmatizer. This research will be helpful in tasks like document classification, contextual advertising, and recommender systems.

References

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 591–600.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, p. 2003, 2003.

[3] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.

[4] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," *arXiv preprint arXiv:1206.3298*, 2012.

[5] N. Kawamae, "Trend analysis model: trend consists of temporal words, topics, and timestamps," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 317–326.

[6] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 424–433.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[8] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," *Journal of Computer and System Sciences*, vol. 61, no. 2, pp. 217–235, 2000.

[9] T. Hofmann, "Probabilistic latent semantic indexing," in *ACM SIGIR Forum*, 2017, vol. 51, no. 2, pp. 211–218.

[10] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Advances in neural information processing systems*, 2005, pp. 1385–1392.

[11] D. Surian, D. Q. Nguyen, G. Kennedy, M. Johnson, E. Coiera, and A. G. Dunn, "Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection," *Journal of medical Internet research*, vol. 18, no. 8, 2016.

[12] D. Ghosh and R. Guha, "What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System," *Cartography and geographic information science*, vol. 40, no. 2, pp. 90–102, 2013.

[13] D. Alvarez-Melis and M. Saveski, "Topic modeling in twitter: Aggregating tweets by conversations," 2016.

[14] R. Lu and Q. Yang, "Trend analysis of news topics on twitter," *International Journal of Machine Learning and Computing*, vol. 2, no. 3, p. 327, 2012.

[15] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 2010, pp. 1155–1158.

[16] J. H. Lau, N. Collier, and T. Baldwin, "Online trend analysis with topic models:# twitter trends detection topic model online," in *Proceedings of COLING 2012*, 2012, pp. 1519–1534.

[17] H. Sha, M. A. Hasan, G. Mohler, and P. J. Brantingham, "Dynamic topic modeling of the COVID-19 Twitter narrative among US governors and cabinet executives," *arXiv preprint arXiv:2004.11692*, 2020.

[18] E. De Santis, A. Martino, and A. Rizzi, "An Infoveillance System for Detecting and Tracking Relevant Topics From Italian Tweets During the COVID-19 Event," *IEEE Access*, vol. 8, pp. 132527–132538, 2020.

[19] L. Singh *et al.*, "A first look at COVID-19 information and misinformation sharing on Twitter," *arXiv preprint arXiv:2003.13907*, 2020.

[20] M. O. Lwin*et al.*, "Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends," *JMIR public health and surveillance*, vol. 6, no. 2, p. e19447, 2020.

[21] C. Ordun, S. Purushotham, and E. Raff, "Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs," *arXiv preprint arXiv:2005.03082*, 2020.

[22] M. Kabir, S. Madria, and others, "CoronaVis: A Real-time COVID-19 Tweets Analyzer," *arXiv preprint arXiv:2004.13932*, 2020.

[23] S. Chaupattnaik, S. S. Nanda, and S. Mohanty, "A suffix stripping algorithm for Odia stemmer," *International Journal of Computational Linguistics and Natural Language Processing*, vol. 1, no. 1, pp. 1–5, 2012.

[24] D. P. Sethi, "Design of lightweight stemmer for Odia derivational suffixes," *Int. Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 12, 2013.

[25] H. Padhy and S. Mohanty, "Designing hybrid approach spell checker for Oriya," *Int. J. Latest Trends Eng. Technol*, vol. 2, no. 4, pp. 156–160, 2013.

[26] I. Jena, S. Chaudhury, H. Chaudhry, and D. M. Sharma, "Developing Oriya morphological analyzer using Lt-toolbox," in *Information Systems for Indian Languages*, Springer, 2011, pp. 124–129.

[27] R. Balabantaray, M. Jena, and S. Mohanty, "Shallow morphology based complex predicates extraction in Oriya," *International Journal of Computer Applications*, vol. 975, p. 8887, 2011.

[28] R. Mohapatra and L. Hembram, "Morph-Synthesizer for Oriya Language-A Computational Approach," *Language In India*, vol. 10, no. 9, 2010.

[29] R. Balabantaray, S. Lenka, and D. Sahoo, "Name Entity Recognizer for Odia using Conditional Random Fields," *Indian Journal of Science and Technology*, vol. 6, no. 4, pp. 4290–4293, 2013.

[30] S. Biswas, S. Mohanty, and S. Mishra, "A hybrid Oriya named entity recognition system: integrating HMM with MaxEnt," in *2009 Second International Conference on Emerging Trends in Engineering & Technology*, 2009, pp. 639–643.

[31] M. K. Jena and S. Mohanty, "Predicting Sensitivity of Local News Articles from Odia Dailies," in *International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making*, 2019, pp. 144–151.

[32] M. K. Jena and S. Mohanty, "Predicting Impact of Odia Newspaper Articles on Public Opinion," in *Progress in Computing, Analytics and Networking*, Springer, 2020, pp. 265–272.

[33] G. Mohanty, P. Mishra, and R. Mamidi, "Annotated Corpus for Sentiment Analysis in Odia Language," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 2788–2795.

[34] W. L. Buntine, "Operations for learning with graphical models," *Journal of artificial intelligence research*, vol. 2, pp. 159–225, 1994.

[35] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 569–577.

[36] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," in *Proceedings of the international working conference on advanced visual interfaces*, 2012, pp. 74–77.

[37] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 2010, pp. 100–108.

Authors' Biography

Swarupananda Bissoyi is currently working as an Asst. Professor in the Department of Computer Application, North Orissa University, Odisha. He has over 11 years of experience in the software industry having worked in Samsung R&D Institute India - Bangalore and Comviva Technologies, Bangalore, and 6 years of academic experience in terms of teaching & research. His research interest includes Data Mining, Recommender Systems, Digital Humanities, and Natural Language processing.

Brojo Kishore Mishra, Ph.D. in Computer Science from Berhampur University, 2012 for his excellent work in the field of Web Mining. He worked in several reputed private Engineering colleges at a different level for more than 15 years. Presently, he is a Professor with the Department of Computer Science and Engineering, GIET University, Gunupur, India, and working as a Chair - R10 SAC for the IEEE Kolkata Section and IEEE Bhubaneswar subsection and Regional Student Coordinator for Computer Society of India Region – IV. He has published more than 30 research papers mainly on peer-reviewed International Journals indexed in Scopus, ESCI, and SCI, and 26 research papers in proceedings of National /International conferences, 34 book chapters, and 09 edited books 02 authored books, 02 patents, and 03 book series. He has successfully guided one Ph.D. research scholar and 04 research scholars are continuing. He served in the capacity of keynote speaker, program chair, proceeding chair, publicity chair, and advisory board members of many international conferences. He is also a life member of ISTE, CSI, and a senior member of IEEE. Also served as a Board of study member for Khallikote Autonomous College, S.B. R. Women Autonomous College, VikramDev Autonomous College.

Raghvendra Kumar is working as Associate Professor in Computer Science and Engineering Department at GIET University, India. He received B. Tech, M.Tech, and Ph.D. in Computer Science and Engineering, India, and Postdoc Fellow from the Institute of Information Technology, Virtual Reality, and Multimedia, Vietnam. He serves as Series Editor Internet of Everything (IOE): Security and Privacy Paradigm, Green Engineering and Technology: Concepts and Applications, publishes by CRC Press, Taylor & Francis Group, USA, and Bio-Medical Engineering: Techniques and Applications, Publishes by Apple Academic Press, CRC Press, Taylor & Francis Group, USA. He also serves as an acquisition editor for Computer Science by Apple Academic Press, CRC Press, Taylor & Francis Group, USA. He has published several research papers in an international journal (SCI/SCIE/ESCI/Scopus) and conferences, including IEEE and Springer. He serves as organizing chair (RICE-2019, 2020), volume Editor (RICE-2018), Keynote speaker, session chair, Co-chair, publicity chair, publication chair, advisory board, Technical program Committee members in many international and national conferences and serve as guest editors in many special issues from reputed journals (Indexation: Scopus, ESCI, SCI). He also published 13 chapters in an edited book published by IGI Global, Springer, and Elsevier. His researches areas are Computer Networks, Data Mining, cloud computing, and Secure Multiparty Computations, Theory of Computer Science, and Design of Algorithms. He authored and Edited 23 computer science books in the Internet of Things, Data Mining, Biomedical Engineering, Big Data, Robotics, and IGI Global Publication, USA, IOS Press Netherland, Springer, Elsevier, CRC Press, USA.