





Article

Modeling Topics in DFA-Based Lemmatized Gujarati Text

Uttam Chauhan ¹, Shruti Shah ¹ , Dharati Shiroya ¹, Dipti Solanki ¹, Zeel Patel ¹, Jitendra Bhatia ^{2,*} ,
Sudeep Tanwar ² , Ravi Sharma ³, Verdes Marina ^{4,*} and Maria Simona Raboaca ^{5,6} 

¹ Department of Computer Engineering, Vishwakarma Government Engineering College, Chandkheda, Ahmedabad 382424, India

² Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad 382481, India

³ Ravi Sharma, Centre for Inter-Disciplinary Research and Innovation, University of Petroleum and Energy Studies, Dehradun 248001, India

⁴ Faculty of Civil Engineering and Building Services, Department of Building Services, Technical University of Gheorghe Asachi, 700050 Iasi, Romania

⁵ Doctoral School, University Politehnica of Bucharest, Splaiul Independentei Street No. 313, 060042 Bucharest, Romania

⁶ National Research and Development Institute for Cryogenic and Isotopic Technologies—ICSI Rm. Vâlcea, Uzinei Street, No. 4, P.O. Box 7 Râureni, 240050 Râmnicu Vâlcea, Romania

* Correspondence: jitendra.bhatia@nirmauni.ac.in (J.B.); marina.verdes@academic.tuiasi.ro (V.M.)

Abstract: Topic modeling is a machine learning algorithm based on statistics that follows unsupervised machine learning techniques for mapping a high-dimensional corpus to a low-dimensional topical subspace, but it could be better. A topic model's topic is expected to be interpretable as a concept, i.e., correspond to human understanding of a topic occurring in texts. While discovering corpus themes, inference constantly uses vocabulary that impacts topic quality due to its size. Inflectional forms are in the corpus. Since words frequently appear in the same sentence and are likely to have a latent topic, practically all topic models rely on co-occurrence signals between various terms in the corpus. The topics get weaker because of the abundance of distinct tokens in languages with extensive inflectional morphology. Lemmatization is often used to preempt this problem. Gujarati is one of the morphologically rich languages, as a word may have several inflectional forms. This paper proposes a deterministic finite automaton (DFA) based lemmatization technique for the Gujarati language to transform lemmas into their root words. The set of topics is then inferred from this lemmatized corpus of Gujarati text. We employ statistical divergence measurements to identify semantically less coherent (overly general) topics. The result shows that the lemmatized Gujarati corpus learns more interpretable and meaningful subjects than unlemmatized text. Finally, results show that lemmatization curtails the size of vocabulary decreases by 16% and the semantic coherence for all three measurements—Log Conditional Probability, Pointwise Mutual Information, and Normalized Pointwise Mutual Information—from -9.39 to -7.49 , -6.79 to -5.18 , and -0.23 to -0.17 , respectively.

Keywords: topic models; Gujarati text lemmatization; Latent Dirichlet Allocation; poor quality topics; overly general topics



Citation: Chauhan, U.; Shah, S.; Shiroya, D.; Solanki, D.; Patel, Z.; Bhatia, J.; Tanwar, S.; Sharma, R.; Marina, V.; Raboaca, M.S. Modeling Topics in DFA-Based Lemmatized Gujarati Text. *Sensors* **2023**, *23*, 2708. <https://doi.org/10.3390/s23052708>

Academic Editors: Chang Choi, Kiho Lim and Gyuhoo Choi

Received: 5 February 2023

Revised: 18 February 2023

Accepted: 24 February 2023

Published: 1 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Topic modeling is statistical modeling for uncovering abstract “topics” hidden in a massive text collection. For example, Latent Dirichlet Allocation (LDA) infers the topics in a text collection [1]. Linguistic field researchers have shown great interest in techniques for discovering a smaller set of word clusters (known as topics) that represents the whole corpus without losing its significance. The set of techniques for modeling topics in domains are Latent Semantic Analysis (LSA) [2], probabilistic Latent Semantic Analysis (pLSA) [3], followed by LDA.

Practitioners have been using topic models to explore the semantic and statistical properties of text corpora. They have successfully applied the technique to a variety of text domains such as scientific and research article corpora [4–7], health and clinical areas [8–11], software domains [12–19], etc. The application of topic models has also been expanded to non-textual data such as (1) a video corpus for person re-identification [20] and human-action recognition [21], (2) an image collection to reorganize images into groups based on their quality [22], and (3) an audio dataset for retrieving audio using the features of audio [23]. Additionally, in various research tasks, a hierarchy of topics has been modeled as opposed to flattened topic modeling [24–29]. Additionally, topic modelers have also explored short texts, such as tweets on Twitter [30–36] or customer reviews [37–39], for discovering hidden thematic structures. The topic model presupposes that each document in the collection is a combination of various topics and that each is a combination of words. A probability distribution across the vocabulary words is subjective. Table 1 illustrates themes derived from a collection of English newspapers.

Table 1. Topics with top 10 words.

Topic 1		Topic 2		Topic 3		Topic 4	
Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.
Ballot	0.073	NYSE	0.104	Gym	0.064	Fire	0.081
Voting	0.071	Predict	0.082	Guideline	0.062	Fundamental	0.079
Poll	0.069	Profitability	0.082	Diet	0.060	Force	0.077
Booth	0.064	NASDAQ	0.073	Fitness	0.060	Galaxy	0.077
Campaign	0.062	Negotiable	0.073	Grains	0.059	Earth	0.077
Election	0.060	Profit	0.073	Growth	0.059	Experimental	0.075
Democracy	0.057	Peak	0.068	Doctor	0.057	Energy	0.069
Leadership	0.053	Portfolio	0.062	Yoga	0.055	Explosion	0.063
Electoral	0.050	Price	0.061	Health	0.055	Star	0.063

One of the topic models' byproducts—topics—can be used either directly in information extraction or as an intermediate output that serves as an input for the subsequent task phase.

Despite numerous extensions worldwide, some areas of LDA still call for more reflection. Preprocessing techniques like stopword removal, stemming, and lemmatization must be created for many languages. Although they can seem like a straightforward component of text summarization, their existence or absence has a significant impact on the output since a thorough evaluation of these preprocessing activities results in more meaningful topics in a shorter period of time. However, the enormous breadth of the vocabulary may come from excluding such stages. The inference procedure consequently requires greater processing resources. Furthermore, less emphasis has been placed on linguistic features like synonyms, polysemy, homonymy, hyponymy, and so forth. These language traits improve the issues' semantic coherence. Also crucial to topic modeling are the preprocessing elements. Language-specific preprocessing is frequently used in NLP research assignments. Instead of getting rid of language-specific stopwords, Schofield et al. suggested topic models preceded by corpus-specific stopwords [40].

1.1. Motivation

Stochastic topic models uncover the latent topical structure, which is smaller in size and easier to understand. However, they need to improve the output at times. The vocabulary size in the text collection of a morphologically rich language increases with the increase in the size of the corpus. It is a fact that topic models transform a mammoth text collection into a manageable topical subspace that is easier to interpret; however, the training phase of LDA may prove itself computationally expensive in the case of the huge size of the vocabulary. This phenomenon is because the statistical inference process continu-

ously refers to the vocabulary. If we reduce the vocabulary size without disturbing the quality of the corpus, the inference process computation cost can be decreased. Moreover, the semantic coherence of topics could also be increased remarkably.

1.2. Contribution of the Paper

The main lines of the contribution process of this paper consist of the following:

- We propose a DFA-based lemmatization approach for Gujarati text.
- We show that lemmatization of Gujarati text reduces lemmas to their base words to curtail the vocabulary size notably.
- The topic can be inferred quicker in a lemmatized Gujarati corpus, resulting in improvement in the interpretability of the discovered topics at the same time.
- The semantic coherence measurement has been performed by three methods to analyze it precisely.
- Additionally, we have used two different measurement methods to show the distance between topics. We proved that meaningful and precise topics fall far from overly general topics. The distance of the meaningful topics from the token distribution of the entire corpus is also larger compared to that for overly general topics.

1.3. Organization of the Paper

The rest of the paper has been organized as follows: Section 2 covers the literature study relevant to the proposed methodology. Then, Section 3 explains the DFA-based approach for lemmatization. Following this, Section 4 briefs about topic inference techniques with their parameters. It depicts some relevant figures of automata. Additionally, a few rules for the first word of the sentence and the rest of the sentences have been shown in tabular format. Next, Section 5 discusses the experimental setup and measurement techniques. Finally, Section 6 displays the experimental findings and their comparison.

1.4. Scope of the Paper

The paper explains the effect of lemmatized text for modeling topics. The technique applies to Gujarati text specifically and to the dataset under study; changes may be needed for it to work more efficiently on another dataset of Gujarati text.

2. Related Work

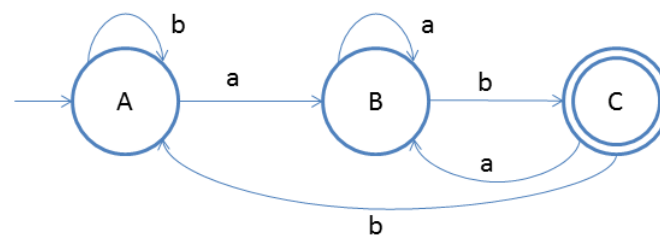
Although preprocessing of the corpus has been considered a very obvious phase, it exhibits challenges when dealing with languages that have a rich morphology. This is because the stemming and lemmatization process differs from one language to another. Most of the topic modeling research tasks target the corpus of English text, as the resources are available for preprocessing. There are several earlier works about learning topic models preceded by stemming or lemmatization. Brahmi et al. modeled the topics in stemmed Arabic text [41]. They achieved two main objectives: first, to extract the stem from the morphemes, and second, to infer topics from Arabic text using LDA. Lu et al. investigated the impact of removing frequent terms from the vocabulary [42]. They measured the computational overhead for different numbers of topics. Designing lemmatization for many languages has captured the attention of linguists. For languages like Hindi [43], academics have created lemmatization of the Indian language, such as Bengali [44] and Tamil [45]. Likewise, Al-Shammari et al. proposed Arabic lemmatization techniques [46] and stemming in another work [47]. Roth et al. also designed an Arabic lemmatizer with feature ranking [48]. The European languages lemmatization approaches include French [49], Polish [50], Finnish [51], Czech [52], German, Latin [53], and Greek [54]. Similarly, the Kazakh language [55], Turkish [56], and Urdu [57,58] have also been considered for lemmatization. Table 2 shows the lemmatization work for the different languages.

Table 2. Lemmatization for different languages.

Author	Language	Application	Pub. Year	Approach	Accuracy	No. of Tokens
[59]	Assamese	Word Sense Disambiguation	2022	Rule-based	82	50,000
[60]	Arabic	Annotation	2018	Dictionary-based	98.6	46,018
[44]	Bengali	Word Sense Disambiguation	2016	Rule-based	96.99	6341
[43]	Hindi	Time Complexity	2013	Rule-based	89.02	2500
[49]	French	Pos Tagging	2010	Rule-based	99.28	350,931
[55]	Kazakh	Information Retrieval	2019	Rule-based	N/A	N/A
[46]	Arabic	Lexem Models	2018	Feature Ranking	N/A	N/A

3. Deterministic Finite Automata (DFA) Based Gujarati Lemmatizer

DFAs, also known as deterministic finite automata, are finite state machines that accept or reject character strings by parsing them through a sequence specific to each string. It is said to be “deterministic” when each string, and thus each state sequence, is distinct. Each input symbol in a DFA moves to the next state that can be predicted as a string of symbols is parsed via DFA. For example, if you want to parse all strings in the alphabet a,b that end with ‘ab,’ then Figure 1 depicts the DFA that accepts only the correct strings.

**Figure 1.** A DFA accepting the strings ends with ‘ab’.

There are two approaches for generating a root word from its inflected word: stemming and lemmatization. Stemming is the method to remove the suffixes/prefixes of the words to get the root words [61]. Lemmatization refers to deriving the root words from the inflected words. A lemma is the dictionary form of the word(s) in the field of morphology or lexicography. To achieve the lemmatized forms of words, one must analyze them morphologically and have the dictionary check for the correct lemma. As a result, the lemmatized word always conveys a proper meaning, while a stemmed word may come out without any meaning. Table 3 explains the difference between stemming and lemmatization. It can be observed that a stemmed word may or may not be the dictionary word, while a lemma must be a dictionary word.

Table 3. Stemming and lemmatization.

Word	Stemming	Lemmatization
Information	Inform	Information
Informative	Inform	Informative
Computers	Comput	Computer
Feet	Feet	Foot

For example, using the continuous bag-of-words model, word embedding applications consider the N (size of windows) surrounding context words to predict the word. Hence, before vocabulary building takes place, the words of the text collection must be preprocessed in terms of stopwords removal, stemming, lemmatizing, etc. This results in the shrinkage of vocabulary size and speeds up the model-building process. Similarly, morphologically, topic modeling in the text of the rich language needs to process a massive vocabulary during the topic formation process. One must apply the lemmatization technique to the corpus to have a reduced vocabulary size. This paper discusses this issue, considering Gujarati (the 26th most widely spoken language in India by the number

of native speakers [62]) for examination. Tables 4 and 5 enlist rules for the lemmatization of inflectional forms. Besides, Figure 2a,b depict rule 1 and rule 2 of Table 4 respectively. Similarly, Figure 3a,b depict rule 3 and rule 4 of Table 4 respectively.

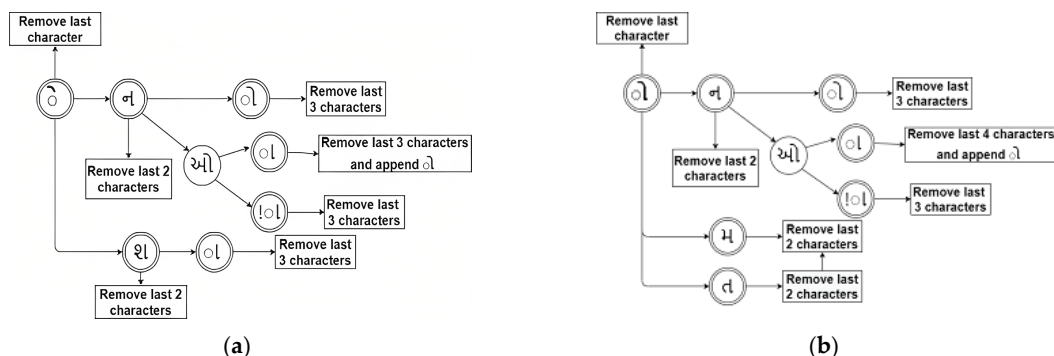


Figure 2. Deterministic finite automata for lemmatization for rules 1 and 2. (a) Rule 1; (b) Rule 2.

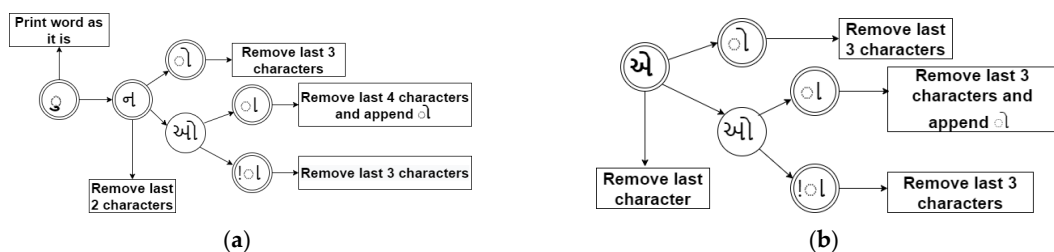


Figure 3. (a) Rule 3; (b) Rule 4. Deterministic finite automata for lemmatization for rules 3 and 4.

Table 4. Part 1: Rules for the first word of the sentence.

Sr. No	Rule Name	How Many Letters / Characters to Check	Letters	What to Delete from Word	What to Add after Deletion	Example
1	Check if the last letter is 'ો'	last 3 characters	'ો', 'ન', 'ો'	last 3 characters	NA	મહાપુરુષોનો = મહાપુરુષ
2	Check if the last letter is 'ો'	last 4 characters	ો', 'ન', 'ઓ', 'ા'	last 4 characters	'ો'	છોકરાઓનો = છોકરો
3	Check if the last letter is 'ો'	last 4 characters	'ો', 'ન', 'ઓ' not 'ા'	last 3 characters	NA	છોકરીઓનો = છોકરી
4	Check if the last letter is 'ો'	last 2 characters	'ો', 'ન'	'ો' and check the remaining word with the words in the n-ending words file. If a match occurs, then print the word; else, remove 'ો', 'ન'	NA	વાહનો = વાહન
5	Check if the last letter is 'ો'	last 2 characters	'ો', 'ન'	last 2 characters	NA	સીતાનો = સીતા
6	Check if the last letter is 'ી'	last 3 characters	'ી', 'ન', 'ી'	last 3 characters	NA	મહાપુરુષોની = મહાપુરુષ
7	Check if the last letter is 'ી'	last 4 characters	'ી', 'ન', 'ઓ', 'ા'	last 4 characters	'ો'	છોકરાઓની = છોકરો
8	Check if the last letter is 'ી'	last 4 characters	'ી', 'ન', 'ઓ' and not 'ા'	last 3 characters	NA	છોકરીઓની = છોકરી
9	Check if the last letter is 'ી'	last 2 characters	'ી', 'ન'	last 2 characters	NA	સીતાની = સીતા
10	Check if the last letter is 'ી'	last 3 characters	'ી', 'થ', 'ો'	last 3 characters	NA	મહાપુરુષોથી = મહાપુરુષ

Table 5. Part 2: Rules for the rest of the words of the sentence.

Sr. No	Rule Name	How Many Letters / Characters to Check	Letters	What to Delete from Word	What to Add after Deletion	Example
1	Check if the last letter is 'ૃ':	last 3 characters	'ૃ', 'ય', '્'	last 3 characters	'વ', 'ૃ'	બન્ધુ = બનવુ
2	Check if the last letter is 'ૃ':	last 4 characters	'લ', 'ે', 'ય', 'ા'	last 4 characters	'વ', 'ૃ'	સંતાડાચેલુ = સંતાડવુ
3	Check if the last letter is 'ૃ':	last 2 characters	'ૃ', 'વ'	last 2 characters	NA	રમવુ = રમ
4	Check if the last letter is 'ૃ':	last 2 characters	'ૃ', 'ત'	last 2 characters	NA	રમતુ = રમ
5	Check if the last letter is 'ૃ':	last 2 characters	'ૃ', 'મ'	last 2 characters	NA	પાંચમુ = પાંચ
6	Check if the last letter is 'ૃ':	last 2 characters	'ૃ', 'શ'	last 3 characters	NA	આવીશુ = આવ
7	Check if the last letter is 'ૃ':	last 3 characters	'ૃ', 'ન', 'ો'	last 3 characters	NA	મહાપુરુષોનુ = મહાપુરુષ
8	Check if the last letter is 'ૃ':	last 4 characters	'ૃ', 'ન', 'ઓ', 'ા'	last 4 characters	'ો'	છોકરાઓનુ = છોકરો
9	Check if the last letter is 'ૃ':	last 4 characters	'ૃ', 'ન', 'ઓ' not 'ા'	last 3 characters	NA	છોકરીઓનુ = છોકરી
10	Check if the last letter is 'ૃ':	last 2 characters	'ૃ', 'ન'	last 2 characters	NA	સીતાનુ = સીતા

In previous work for normalizing the word forms in Gujarati, the stemming approach has received attention from linguistic researchers. Patel et al. prepared a suffix list and incorporated it into the stemming process. They targeted to get rid of only suffixes of the inflectional words. Likewise, Suba et al. proposed an extended version of stemmer, which is lightweight for suffix removal and a heavyweight rule-based stemmer [63]. Ameta et al. also suggested a similar kind of lightweight stemmer [64]. Aswani et al. offered a morphological study for inflectional forms of the Hindi language, which was extended to the Gujarati language [65]. In all previous approaches, authors have focused on suffix removal to reduce the word to its root. We perform mainly three operations for transforming inflectional forms to a lemma: removing suffixes, appending suffixes, and removal followed by appending of characters.

For the Gujarati language, lemmatization is more accurate in transforming the inflected word into the root word. It involves not only the removal of suffixes but also appending some pattern or characters to the inflected word or the operations one after another. There is a remarkable set of research tasks stemming from the Gujarati language, such as hybrid stemmer of Gujarati [66], Gujarati lightweight stemmer [64], and rule-based stemmer [63]. However, research on lemmatization has not gained much consideration comparatively.

Apart from the research mentioned above, there is hardly any work found in the literature that directly addresses the problems of lemmatization in the Gujarati language. In this article, we propose an automata-based approach for lemmatizing Gujarati text. Besides the automata-based approach, list-based (known as rule-based) and hash-based approaches have also been explored for lemmatization across various languages. We aim to design an automata-based lemmatizer that can transform different inflectional forms to their root word with less computational complexity than the list-based approach.

The different inflectional words are generated by applying one or more transformation rules to their respective dictionary form or lemma. An inflectional word can be formed by appending a valid suffix to its lemma ('મહાપુરુષ' can form the lemma 'મહાપુરુષોથી' by appending the suffix 'ઓથી'). Moreover, sometimes removal followed by addition may result in the formation of a valid inflectional form ('છોકરો' can form the lemma 'છોકરાઓનુ') by appending the suffix 'નુ') or sometimes none of the mentioned cases. Lemmas can be

derived by reversing the corresponding process of generation of inflectional forms from their root words.

4. Latent Dirichlet Allocation (LDA)

In Latent Dirichlet Allocation, each document in the corpus of M document is modeled as a multinomial distribution of K hidden topics, and each topic is a multinomial distribution of the vocabulary of V words. Topic modeler inputs the number of topics, K . The document-topic distribution θ_d is drawn from Dirichlet distribution $\text{Dir}[\alpha]$, where α is a hyperparameter vector variable with the value $(\alpha_1, \alpha_2, \dots, \alpha_k)$, which can be estimated. In the same way, the topic-word distribution, ϕ_k , is drawn from the Dirichlet distribution $\text{Dir}[\beta]$.

The Latent Dirichlet Allocation can be represented graphically by the Bayesian network. The plate notation of LDA has been depicted in Figure 4. The node represents the random variable, and the edge represents the influence of one variable on another. The complete document generative process θ_d and ϕ_k has been shown in Algorithm 1. For the n th word of document d , a topic assignment $Z_{n,d}$ is drawn from θ_d , and a word identity is drawn from the corresponding topic, $\phi_{W|Z_d}$. Henceforth, the whole generative process is given by

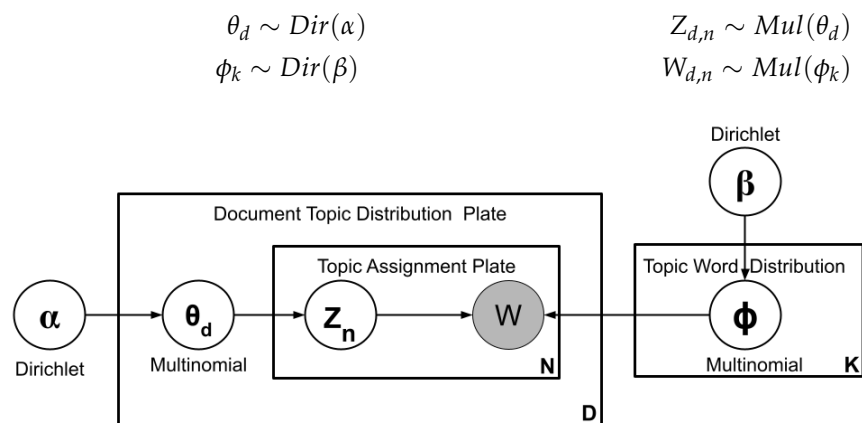


Figure 4. Plate notation for LDA generative algorithm [1].

Algorithm 1: Generative algorithm for LDA.

```

Input: Dataset, K topics, Hyperparameters  $\alpha$  and  $\beta$ 
Output: Topic files, Topic word distribution, Document topic distribution
for All topics  $k \in [1, K]$  do
  // sample the probability distribution of words for each topic
  sample mixture components  $\vec{\phi}_k \sim \text{Dir}(\vec{\beta})$ ;
end
for all documents  $m \in [1, M]$  do
  // proportion of topics for each document
  sample mixture proportion  $\vec{\theta}_m \sim \text{Dir}(\vec{\alpha})$ ;
  // Length of documents in the corpus is normally distributed
  sample document length  $N_m \sim \text{Poisson}(\zeta)$ ;
  for all words  $n \in [1, N_m]$  in document  $m$  do
    // assign the topic to each word
    sample topic index  $Z_{m,n} \sim \text{Mult}(\vec{\theta}_m)$ ;
    // identify the word identity from a probability distribution
    of words
    sample term for words  $W_{m,n} \sim \text{Mult}(\phi_{Z_{m,n}})$ ;
  end
end

```

5. Experimental Setup

We used the TDIL dataset to assess our lemmatization approach's efficiency. The dataset was made available with Part-of-Speech (PoS) tagging, but we wanted to consider something other than PoS tagging, as it is optional. Therefore, we filtered the PoS tags to get a compatible dataset for our experimental purposes. We experimented with evaluating two metrics: the improvement in the semantic coherence of topics and the success rate of lemmatization. Initially, the lemmatization process was evaluated in terms of accuracy. We provided the dataset and measured how many words were correctly lemmatized and how many were lemmatized incorrectly. We also considered words that did not need any lemmatization separately. In other words, no rule should apply to such words.

5.1. Preprocessing and Vocabulary Size: An Analysis

Although preprocessing of the dataset is a self-explanatory part of the text analysis, it may uncover the detailed statistical properties of the corpus under study. Moreover, preprocessing is also language-dependent. Hence, a specific set of preprocessing steps is required for a given language to know the values of some critical parameters such as the number of tokens in the corpus, vocabulary size, type-to-token ratio (TTR), etc. In TTR, tokens denote the total number of words in a corpus regardless of how often they are repeated. The term "type" concerns the number of distinct words in a corpus. We performed preprocessing for the dataset under study. The preprocessor component comprises several steps, as depicted in Table 6. As shown in Table 6, stopwords removal remarkably reduced the size of the dataset in terms of the total number of tokens and discrete words. We compiled a list of more than 800 Gujarati language stopwords to eliminate them.

Table 6. Tokens, vocabulary, and TTR.

Preprocessing Steps	No. of Tokens	Vocabulary Size	% of Tokens in Vocabulary	TTR
After tokenization	1,167,630	89,696	7.681885529	0.077
After stopwords removal	870,521	89,003	10.22410717	0.102
After punctuation removal	746,292	889.87	11.92388502	0.119
Alphanumeric to alphabetic word	746,292	86,271	11.5599524	0.116
After single-letter word removal	620,133	86,098	13.8837959	0.139
After lemmatization	620,133	50,043	8.069720528	0.081

The punctuation marks and alphanumeric words are typical cases in many languages. However, Gujarati text might contain Gujarati and English digits blended in the text, so we have taken care to remove the digits of both languages. We also found several words in alphanumeric form, so we transformed them into alphabetic forms by removing mixed digits from such words. However, they are rare in number, and this did not decrease the number of tokens and vocabulary size. The next step is particular to Gujarati text, as the Gujarati corpus contains single-letter words. Table 7 depicts the most frequent single-letter words in the Gujarati language. These words do not contribute to topic modeling and are not part of the stopwords. The crucial part is that we performed lemmatization on the resultant corpus. It reduced the inflectional forms to their root words, known as the lemma.

Table 7. Single letter words.

Word	Probability	Word	Probability
'કે' (Kē / Whether)	0.003833333	'જો' (Jō / If)	0.000750000
'છે' (Chhē / Is)	0.025166667	'જા' (Ja / Only)	0.002500000
'જે' (Jē / Whom)	0.001083333	'ના' (Na / No)	0.001000000
'તે' (Tē / That)	0.001666667	'બે' (Bē / Two)	0.001166667
'એ' (Ē / That)	0.000833333	'તો' (Tō / Then)	0.001833333
'આ' (Ā / This)	0.004250000	'હુ' (Hu / I)	0.000416667
'છો' (Chho / Is)	0.000833333	'શ્રી' (Shree / Mr.)	0.001583333

We have achieved a remarkable reduction in the vocabulary size, as shown in Figure 5. Moreover, after each preprocessing step, one can observe a notable decrease in the number of tokens. Most importantly, the lemmatization left the vocabulary size at 8.07% of the total number of tokens. The vocabulary size was 7.29% before any preprocessing action, but the number of tokens columns was very high. That itself could lead to the heavy computation of inference of the topic. Meanwhile, there is a negligible reduction in vocabulary after the removal of digits and the transformation of alphanumeric words to alphabetic.

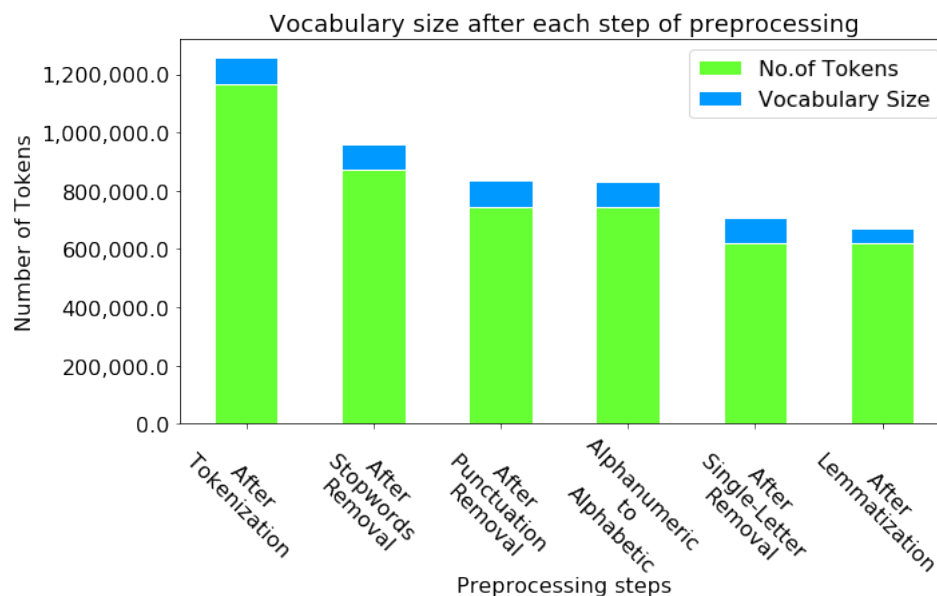


Figure 5. Size of vocabulary.

As shown in Figure 5, although the lemmatization process does not reduce the number of tokens, it reduces vocabulary size by 58% because instead of eliminating the tokens from the collection, it transforms them to their respective lemma form. Similarly, there were several alphanumeric words in the corpus. We removed the blended digits from those alphanumeric words, leaving behind the alphabetic words. The alphanumeric words might occur due to typing errors. However, we did not perform any tasks for removing the words that occur less frequently than some number N ; for example, when N is 3. Other authors have removed words with some lower and upper bounds in frequency in most information retrieval research. For example, the words that occur fewer than three times and more than 500 times are to be removed. In most cases, these words get removed in one of the previous preprocessing steps, such as stopword elimination or single-letter word removal.

5.2. Evaluation of the Proposed Lemmatizer Approach

To evaluate the technique, we prepared a lemma-annotated dataset constructed from the dataset itself. This lemma-annotated dataset was constructed with the help of words found frequently in articles based on terms that appear in all four categories of the dataset. They are dictionary words. As mentioned earlier, we did not include single-letter words, stopwords, or punctuation marks in the lemma-annotated dataset, as they have been removed from the preprocessed dataset. We took 2000 lemmas for the experiment, achieved by stratified sampling. On applying the lemmatizer, we achieved 78.6% accuracy. The outcome of the experiment is regardless of Part-of-Speech tagging.

5.3. Overly General Topics

Several decision criteria could identify overly general topics. A topic model may consist of different categories of overly general topics. An extensive topic needs to make sense

thematically. These themes include a collection of words that have no semantic connection to one another, to put it simply. For example, a topic might cover a significant fraction of the words in the vocabulary. Table 8 depicts an overly general topic, which comprises 11% of vocabulary words. Such topics are very general and do not convey any specific concept. On the other hand, an interpretable topic comprises semantically relevant words. Therefore, one can find some meaning in it, as shown in Table 9. A meaningful topic contains a tiny fraction of the words in the vocabulary, such as 1% to 2%; on the other hand, a few topics may be very common, as they are present in many documents. Furthermore, uninterpretable topics can also be identified by the number of tokens assigned to the top-N words of the topic. Therefore, the word length, i.e., the average number of characters present in the words of the topic, plays a significant part in determining the interpretability of the topic.

Table 8. Global topic or overly general topic.

Word	Frequency	Word	Frequency
ટેક્સ (Tēksa/Tax)	231	જાહેર (Jāhēra/Public)	67
વરસાદ (Varasāda/Rain)	191	પ્રોજેક્ટ (Prōjēkṭa/Project)	57
ગુજરાત (Gujarāta/Gujarat)	189	રકમ (Rakama/Amount)	46
જાહેર (Jāhēra/Public)	182	જમીન (Jamin/Soil)	45
સરકાર (Sarakāra/Government)	170	યોગ (Yōga/Yoga)	39
યોગ (Yōga/Yoga)	147	ગુજરાતમાં (Gujarātamām/In Gujarat)	35
શરૂ (Śarū/Start)	138	પ્લાન (Plan/Plan)	31
ભારતીય (Bhāratīya/Indian)	136	વર્ષ (Varsē/Year)	30
ભારત (Bhārata/India)	126	શક્તિ (Śakti/Power)	27
બુલેટ (Bulēṭa/Bullet)	121	એફઆઈઆઈ (Ēpha'ā'ī'ā'ī/FII)	25
પાણી (Pāṇī/Water)	118	સમયસર (Samaysara/On time)	25
અમદાવાદ (Amadāvāda/Ahmedabad)	114	મહત્વ (Mahtava/Importance)	24
પ્રવેશ (Pravēśa/Entry)	112	વિધુર (Vidhura/Widower)	19
તલાક (Talāka/Divorce)	112	મુંબઈ (Mumbāi/Mumbai)	18
સ્માર્ટફોન (Smārtaphōna/Smartphone)	108
નિર્ણય (Nirṇaya/Decision)	107
બાહુબલી (Bāhubalī/Bahubali)	106

Table 9. Interpretable topica.

Word	Frequency	Word	Frequency
ધર્મ (Dharma/Religion)	86	સાક્ષાત (Sāksāta/Confirmed)	18
આનંદ (Ānanda/Happiness)	41	પૂજાપાઠ (Pūjāpāṭha/Worship)	18
ઈશ્વર (Īśvara/God)	37	જાગૃતિ (Jāgṛti/Awareness)	18
યહુદી (Yahudī/Jew)	32	સાંપ્રદાયિક (Sāmpradāyika/Sectarian)	18
કર્મકાંડ (Karmakāṇḍa/Ritual)	22	પ્રેમ (Prēma/Love)	18
નેતિક (Naitika/Moral)	21	ખ્રિસ્તી (Khrīstī/Christian)	16
શ્રદ્ધા (Śrad'āhā/Devotion)	21	ઈસ્લામ (Islāma/Islam)	16
આધ્યાત્મિક (Ādhyātmika/Spiritual)	20	જીવન (Jīvana/Life)	9
ધાર્મિક (Dhārmika/Religious)	20	પ્રત્યે (Pratyē/Towards)	8
મુલ્યો (Mulyō/Values)	19	માણસ (Mānasa/Human)	8

Distance from a Global Corpus-Level Topic

A topic is a probability distribution of words in the vocabulary of a corpus. Each word in the corpus follows with a specific probability. When topics are inferred from the corpus, the words in the topic also carry the probability values. Here topics are soft clusters so that a word may appear in more than one topic with different probability values. Ideally, words forming the topic are semantically relevant to one other.

5.4. Semantic Coherence Measurement Methods

5.4.1. Pointwise Mutual Information (PMI)

To calculate the collocation, PMI might be used. First, though, it serves as a statistical indicator of the proximity of two words. To track the co-occurrence, we changed the sliding window's word count to 5. The PMI of each set of provided word pairs is then determined by computing $(word_1, word_2)$. The PMI of any two terms in a topic model has calculated the difference between the likelihood of their co-occurrence given their joint probability distribution and their discrete distributions, assuming that events are unrelated to one another [67–69]. It can be written mathematically, as shown.

$$PMI(word_1, word_2) = \log \frac{P(word_1, word_2)}{P(word_1)P(word_2)} \quad (1)$$

The word order does not affect the PMI score for that pair of words. The measurements for PMI $(word_1, word_2)$ and PMI remain symmetric $(word_2, word_1)$. The explanation is that since documents are viewed as a “bag of words” (BOW), the sequence in which words appear does not matter. The word order should be emphasized in the reference corpus too. Both positive and negative values could be assigned to the PMI score. If the PMI value is zero, the words are considered to have no relationship with the reference corpus. On the other hand, PMI is greatest when there is a close relationship between the terms.

5.4.2. Normalized Pointwise Mutual Information (NPMI)

The extension of the PMI method is Normalized PMI. It is similar to PMI except that the score of NPMI takes a value between $[-1, +1]$, in which -1 conveys no occurrence together, 0 indicates independence, and 1 indicates complete co-occurrence [68,69].

$$NPMI(word_1, word_2) = \frac{PMI(word_1, word_2)}{-\log P(word_1, word_2)} \quad (2)$$

5.4.3. Log Conditional Probability (LCP)

Log conditional probability (LCP) is one-sided, while PMI is a symmetric coherence measurement.

$$LCP(word_1, word_2) = \log \frac{P(word_1, word_2)}{P(word_2)} \quad (3)$$

5.5. Distance Measurement Methods

The divergence of the topics from some predefined, overly general topic types has been measured. There are two types of metrics for cluster analysis. Supervised evaluation metrics use the labeled samples. On the other hand, unsupervised evaluation does not check the accuracy of the learned model. In this paper, several divergence measures are used to check the efficacy of the proposed techniques.

5.5.1. Hellinger Distance

Hellinger distance between two discrete probability distributions P and Q can be defined as,

$$HD(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^k (\sqrt{p_i} - \sqrt{q_i})^2} \quad (4)$$

where $P = (p_1, p_2, \dots, p_k)$ and $Q = (q_1, q_2, \dots, q_k)$.

5.5.2. Jaccard Distance

The Jaccard similarity measures the similarity between finite sample sets. It is the intersection of sets divided by the union of sample sets. Here, cardinality represents the number of elements in a set, denoted by $|a|$. Suppose you want to find Jaccard's similarity

between two sets, a and b. It is the ratio of the cardinality of $a \cup b$ and $a \cap b$.

$$J(a, b) = |\text{Intersection}(a, b)| / |\text{Union}(a, b)| \quad (5)$$

Although it seems very simple, it applies to the topic modeling. First, it fetches out the standard terms between two topics and the entire distinct terms. Then, it takes the ratio of common and distinct terms; the results serve as the Jaccard similarity. Finally, by taking the complement, likewise in cosine, the Jaccard distance can be measured.

6. Results

6.1. Distance Measurement from Global Topic

The distance measurement mentioned above is used to analyze the quality of topic models. We framed the global topic as the frequency distribution of words in the vocabulary. We used the Hellinger distance to compute the distance of topics from the global topic. The distance between the topic model and the topic model with lemmatized terms has been compared. An experiment examined the effect of lemmatization on vocabulary size and inferred topic quality. We inferred 100 topics by iterating 500 times over the corpus, document by document, and word by word. The distance of topics from the global topic, the list of words in the vocabulary, and the corresponding frequency have been computed using the Hellinger and Jaccard distance measurement techniques.

Table 10 shows the average of 100 topics inferred from the corpus before and after lemmatization. The test outcomes found that the distance of topics modeled over the lemmatized corpus is more than that of the unlemmatized corpus. This is because interpretability and semantic coherence are the parameters correlated with a distance of topics from the predefined global topic. Therefore, it can be concluded that topics learned through the lemmatized text are more significant than topics learned through unlemmatized text.

Table 10. The distance between topics for the unlemmatized and lemmatized corpus.

No. of Tokens	Vocabulary	Inference Time (in Seconds)	Distance Measurement			
			Unlemmatized		Lemmatized	
			Hellinger	Jaccard	Hellinger	Jaccard
604,389	85,463	33.14	0.476	0.970	0.491	0.993
561,648	42,722	29.63	0.495	0.968	0.546	0.998
531,870	27,533	26.77	0.481	0.982	0.520	0.996
512,085	21,238	22.15	0.489	0.982	0.517	0.999
496,373	17,310	18.92	0.495	0.982	0.528	1.000
483,108	14,657	16.55	0.492	0.983	0.526	0.999

The distance of the topic model before performing the lemmatization of words was compared with lemmatized corpus topic model. Table 10 comprises the experimental outcomes. The distance of lemmatized topics was increased; specifically, the Hellinger distance increased by 3% to 5% and the Jaccard distance by more than 2%. It can be inferred that modeling topics in lemmatized text made them more interpretable and meaningful. Moreover, the distance got wider with the shrinkage of the vocabulary size.

Comparing individual distance values, Figure 6a,b depicts the Hellinger and Jaccard distance of 10 topics from the very general topic, respectively. Each topic of lemmatized text falls farther from the global topic than the topics discovered from the unlemmatized text in both cases. The Hellinger distance increased from 1% to 9%, while the Jaccard distance showed a distance difference within the range of 1% to 3%.

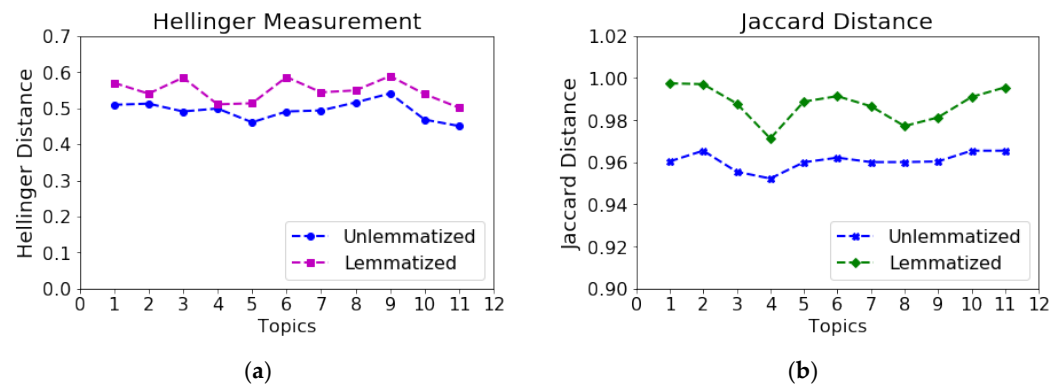


Figure 6. Distance comparison between unlemmatized and lemmatized topics for first 10 topics. (a) Hellinger distance measurement; (b) Jaccard distance measurement.

6.2. The Semantic Coherence Score

We used the Pointwise Mutual Information (PMI), Normalized PMI (NPMI), and Log Conditional Probability methods for the semantic coherence measurement. The semantic coherence of the topics showed improvement after the lemmatization process. All three methods found an increase in the coherence score. In addition to distance measurements, the semantic coherence scores also support that topic models become more interpretable if topics are modeled in the lemmatized text. The coherence value increases as well with the reduction in the size of the vocabulary. The semantic coherence was enhanced up to 3% with LCP and PMI methods and up to 6% for NPMI. Although the topic models found a slight improvement in the semantic coherence values with a decrease in the vocabulary size, the inference time decreased remarkably. Hence, topics learned from lemmatized text are more meaningful than those from unlemmatized text.

We computed the semantic coherence value for 10 topics individually for each technique described. Figure 7a–c depicts the comparison of the coherence value of the unlemmatized text topic model with the lemmatized text topic model. The coherence improvement increased within the range of 2% to 11% for LCP, 2% to 9% for PMI, and up to 1% to 3% for NPMI, while the overall enhancement is lower compared to the 10 topics. However, several topics did not improve semantically. Moreover, the coherence values decreased for a few topics as well. These points caused the average enhancement to have lower values than the first 10 topics.

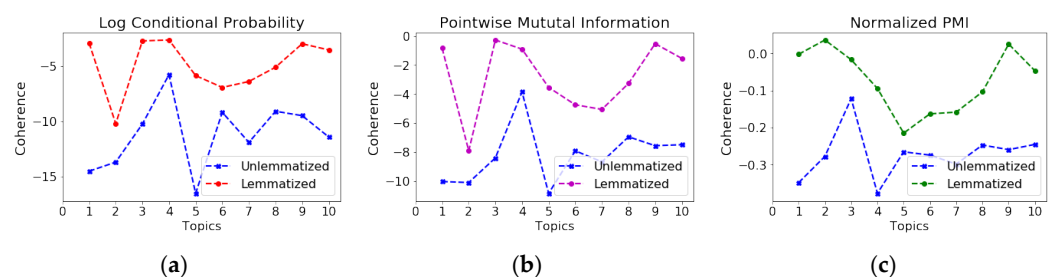


Figure 7. Semantic coherence value comparison between unlemmatized and lemmatized topics for the first 10 topics. (a) Log Conditional Probabilities; (b) Pointwise Mutual Information; (c) Normalized PMI.

7. Conclusions

In this paper, we have proposed a DFA-based lemmatization approach for Gujarati text. The proposed method comprises more than 59 rules for lemmatizing Gujarati text. We have managed to lemmatize 83% of words correctly. To examine the effect of the proposed lemmatization approach on text analysis, we applied LDA for inferring topics from

a text corpus. We showed that the vocabulary size was reduced drastically when lemmatization was involved, although the number of tokens did not decrease. The experimental outcomes revealed that the interpretability of topics increased when the corpus was lemmatized. The topics became more precise and meaningful. This finding was supported by the Hellinger distance and Jaccard distance. Moreover, the semantic coherence measurements supported improving the quality of topics. Our three techniques, PMI, NPMI, and LCP, reported an increase in the coherence value. Furthermore, topics were found to be more specialized when they were modeled from the lemmatized corpus. Moreover, the semantic association among the topics' words has also been enhanced. A generalized approach can be developed for any text corpus in the future. For example, a set of rules can achieve lemmatization for news articles, discussion forums, textbooks, novels, social media text domains, etc.

Author Contributions: Conceptualization: U.C., S.S., D.S. (Dharati Shiroya), D.S. (Dipti Solanki), J.B. and S.T.; Writing—original draft preparation: Z.P., V.M., M.S.R., R.S. and S.T.; Methodology: S.T., V.M., U.C., D.S. (Dharati Shiroya) and J.B.; Writing—review and editing: S.T., D.S. (Dipti Solanki), Z.P., U.C. and V.M.; Investigation: M.S.R., S.T., V.M., J.B. and U.C.; Supervision: S.T., S.S., R.S. and V.M.; Visualization; S.S., D.S. (Dharati Shiroya), D.S. (Dipti Solanki) and M.S.R.; Software: J.B., U.C., S.T. and V.M. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was partially supported by UEFISCDI Romania and MCI through BEIA projects AutoDecS, SOLID-B5G, T4ME2, DISAVIT, PIMEO-AI, AISTOR, MULTI-AI, ADRIATIC, Hydro3D, PREVENTION, DAFCC, EREMI, ADCATER, MUSEION, FinSESCo, iPREMAS, IPSUS, U-GARDEN, CREATE and by European Union's Horizon Europe research and innovation program under grant agreements No. 101037866 (ADMA TranS4MErs). This work is supported by Ministry of Research, Innovation, Digitization from Romania by the National Plan of R & D, Project PN 19 11, Subprogram 1.1. Institutional performance—Projects to finance excellence in RDI, Contract No. 19PFE/30.12.2021 and a grant of the National Center for Hydrogen and Fuel Cells (CNHPC)—Installations and Special Objectives of National Interest (IOSIN).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No data available to carry out this research.

Acknowledgments: The authors want to thank the "Technology Development for Indian Languages (TDIL) Program under the Ministry of Electronics and Information Technology (MeitY), Govt of India," for providing us the Text Corpora, namely "Hindi-Gujarati Parallel Chunked Text Corpus ILCI-II" for research and experimental purposes <https://npl.t.in/demo/text-corpus/hin-guj-parallel-chunked-text-corpus> (accessed 1 March 2020).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. [[CrossRef](#)]
2. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391. [[CrossRef](#)]
3. Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 30 July–1 August 1999; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1999; pp. 289–296. [[CrossRef](#)]
4. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [[CrossRef](#)] [[PubMed](#)]
5. Yau, C.K.; Porter, A.; Newman, N.; Suominen, A. Clustering scientific documents with topic modeling. *Scientometrics* **2014**, *100*, 767–786. [[CrossRef](#)]
6. Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; Smyth, P. The author-topic model for authors and documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, Banff, AB, Canada, 7–11 July 2004; AUAI Press: Arlington, VA, USA, 2004; pp. 487–494. [[CrossRef](#)]

7. Steyvers, M.; Smyth, P.; Rosen-Zvi, M.; Griffiths, T. Probabilistic author-topic models for information discovery. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; ACM: New York, NY, USA, 2004; pp. 306–315.
8. Lu, H.M.; Wei, C.P.; Hsiao, F.Y. Modeling healthcare data using multiple-channel latent Dirichlet allocation. *J. Biomed. Inform.* **2016**, *60*, 210–223. [[CrossRef](#)]
9. Paul, M.J.; Dredze, M. Discovering health topics in social media using topic models. *PLoS ONE* **2014**, *9*, e103408. [[CrossRef](#)]
10. Kayi, E.S.; Yadav, K.; Chamberlain, J.M.; Choi, H.A. Topic Modeling for Classification of Clinical Reports. *arXiv* **2017**, arXiv:1706.06177.
11. Yao, L.; Zhang, Y.; Wei, B.; Wang, W.; Zhang, Y.; Ren, X.; Bian, Y. Discovering treatment pattern in Traditional Chinese Medicine clinical cases by exploiting supervised topic model and domain knowledge. *J. Biomed. Inform.* **2015**, *58*, 260–267. [[CrossRef](#)]
12. Asuncion, H.U.; Asuncion, A.U.; Taylor, R.N. Software traceability with topic modeling. In Proceedings of the 2010 ACM/IEEE 32nd International Conference on Software Engineering, Cape Town, South Africa, 2–8 May 2010; Volume 1, pp. 95–104.
13. Chen, T.H.; Shang, W.; Nagappan, M.; Hassan, A.E.; Thomas, S.W. Topic-based software defect explanation. *J. Syst. Softw.* **2017**, *129*, 79–106. [[CrossRef](#)]
14. Corley, C.S.; Damevski, K.; Kraft, N.A. Changeset-based topic modeling of software repositories. *IEEE Trans. Softw. Eng.* **2018**, *46*, 1068–1080. [[CrossRef](#)]
15. Lukins, S.K.; Kraft, N.A.; Etkorn, L.H. Bug localization using latent dirichlet allocation. *Inf. Softw. Technol.* **2010**, *52*, 972–990. [[CrossRef](#)]
16. Řehůřek, R.; Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*; ELRA: Valletta, Malta, 2010; pp. 45–50.
17. Sun, X.; Li, B.; Leung, H.; Li, B.; Li, Y. Msr4sm: Using topic models to effectively mining software repositories for software maintenance tasks. *Inf. Softw. Technol.* **2015**, *66*, 1–12. [[CrossRef](#)]
18. Thomas, S.W.; Adams, B.; Hassan, A.E.; Blostein, D. Studying software evolution using topic models. *Sci. Comput. Program.* **2014**, *80*, 457–479. [[CrossRef](#)]
19. Tian, K.; Revelle, M.; Poshyvanyk, D. Using latent dirichlet allocation for automatic categorization of software. In Proceedings of the 2009 6th IEEE International Working Conference on Mining Software Repositories, Vancouver, BC, Canada, 16–17 May 2009; pp. 163–166.
20. Vretos, N.; Nikolaidis, N.; Pitas, I. Video fingerprinting using Latent Dirichlet Allocation and facial images. *Pattern Recognit.* **2012**, *45*, 2489–2498. [[CrossRef](#)]
21. Fernandez-Beltran, R.; Pla, F. Incremental probabilistic Latent Semantic Analysis for video retrieval. *Image Vis. Comput.* **2015**, *38*, 1–12. [[CrossRef](#)]
22. Yuan, B.; Gao, X.; Niu, Z.; Tian, Q. Discovering Latent Topics by Gaussian Latent Dirichlet Allocation and Spectral Clustering. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2019**, *15*, 25. [[CrossRef](#)]
23. Hu, P.; Liu, W.; Jiang, W.; Yang, Z. Latent topic model for audio retrieval. *Pattern Recognit.* **2014**, *47*, 1138–1143. [[CrossRef](#)]
24. Gao, N.; Gao, L.; He, Y.; Wang, H.; Sun, Q. Topic detection based on group average hierarchical clustering. In Proceedings of the 2013 International Conference on Advanced Cloud and Big Data, Nanjing, China, 13–15 December 2013; pp. 88–92. [[CrossRef](#)]
25. Kim, D.; Oh, A. Hierarchical Dirichlet scaling process. In Proceedings of the International Conference on Machine Learning, Beijing, China, 22–24 June 2014; pp. 973–981.
26. Li, W.; Yin, J.; Chen, H. Supervised Topic Modeling Using Hierarchical Dirichlet Process-Based Inverse Regression: Experiments on E-Commerce Applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *30*, 1192–1205. [[CrossRef](#)]
27. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Sharing clusters among related groups: Hierarchical Dirichlet processes. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 5–8 December 2005; pp. 1385–1392.
28. Yang, S.; Yuan, C.; Hu, W.; Ding, X. A hierarchical model based on latent dirichlet allocation for action recognition. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 2613–2618. [[CrossRef](#)]
29. Zhu, W.; Zhang, L.; Bian, Q. A hierarchical latent topic model based on sparse coding. *Neurocomputing* **2012**, *76*, 28–35. [[CrossRef](#)]
30. Fang, A.; Macdonald, C.; Ounis, I.; Habel, P. Topics in tweets: A user study of topic coherence metrics for Twitter data. In Proceedings of the European Conference on Information Retrieval, Padua, Italy, 20 March 2016; Springer: Cham, Switzerland, 2016; pp. 492–504. [[CrossRef](#)]
31. Weng, J.; Lim, E.P.; Jiang, J.; He, Q. Twitterrank: Finding topic-sensitive influential twitterers. In Proceedings of the Third ACM International Conference on Web Search and Data Mining, New York, NY, USA, 3–6 February 2010; ACM: New York, NY, USA, 2010; pp. 261–270.
32. Bhattacharya, P.; Zafar, M.B.; Ganguly, N.; Ghosh, S.; Gummadi, K.P. Inferring user interests in the twitter social network. In Proceedings of the 8th ACM Conference on Recommender Systems, Foster City, CA, USA, 6–10 October 2014; ACM: New York, NY, USA, 2014; pp. 357–360.
33. Cordeiro, M. Twitter event detection: combining wavelet analysis and topic inference summarization. In Proceedings of the Doctoral Symposium on Informatics Engineering; Faculdade de Engenharia da Universidade do Porto: Porto, Portugal, 2012; pp. 11–16.

34. Kim, Y.; Shim, K. TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Inf. Syst.* **2014**, *42*, 59–77. [[CrossRef](#)]
35. Lansley, G.; Longley, P.A. The geography of Twitter topics in London. *Comput. Environ. Urban Syst.* **2016**, *58*, 85–96. [[CrossRef](#)]
36. Ren, Y.; Wang, R.; Ji, D. A topic-enhanced word embedding for twitter sentiment classification. *Inf. Sci.* **2016**, *369*, 188–198. [[CrossRef](#)]
37. Ma, B.; Zhang, D.; Yan, Z.; Kim, T. An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews. *J. Electron. Commer. Res.* **2013**, *14*, 304. [[CrossRef](#)]
38. Hashimoto, K.; Kontonatsios, G.; Miwa, M.; Ananiadou, S. Topic detection using paragraph vectors to support active learning in systematic reviews. *J. Biomed. Inform.* **2016**, *62*, 59–65.
39. Kim, S.; Zhang, J.; Chen, Z.; Oh, A.; Liu, S. A hierarchical aspect-sentiment model for online reviews. *Proc. Aaai Conf. Artif. Intell.* **2013**, *27*, 526–533.
40. Schofield, A.; Magnusson, M.; Mimno, D. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Short Papers*; Association for Computational Linguistics: Valencia, Spain, 2017; Volume 2, pp. 432–436.
41. Brahmi, A.; Ech-Cherif, A.; Benyettou, A. Arabic texts analysis for topic modeling evaluation. *Inf. Retr.* **2012**, *15*, 33–53. [[CrossRef](#)]
42. Lu, K.; Cai, X.; Ajiferuke, I.; Wolfram, D. Vocabulary size and its effect on topic representation. *Inf. Process. Manag.* **2017**, *53*, 653–665. [[CrossRef](#)]
43. Paul, S.; Tandon, M.; Joshi, N.; Mathur, I. Design of a rule based Hindi lemmatizer. In *Proceedings of Third International Workshop on Artificial Intelligence, Soft Computing and Applications, Chennai, India, 27 July 2013*; AIRCC Publishing Corporation: Tamil Nadu, India, 2013; pp. 67–74.
44. Chakrabarty, A.; Garain, U. Benlem (A bengali lemmatizer) and its role in WSD. *ACM Trans. Asian-Low-Resour. Lang. Inf. Process. (TALLIP)* **2016**, *15*, 1–18. [[CrossRef](#)]
45. Kumar, A.M.; Soman, K. AMRITA_CEN@ FIRE-2014: Morpheme Extraction and Lemmatization for Tamil using Machine Learning. In *Proceedings of the Forum for Information Retrieval Evaluation, Bangalore, India, 5–7 December 2014*; Association for Computing Machinery: New York, NY, USA, 2014; pp. 112–120.
46. Al-Shammari, E.; Lin, J. A novel Arabic lemmatization algorithm. In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data, Association for Computing Machinery, New York, NY, USA, 24 July 2008*; pp. 113–118.
47. Al-Shammari, E.T.; Lin, J. Towards an error-free Arabic stemming. In *Proceedings of the 2nd ACM Workshop on Improving non English Web Searching, Napa Valley, CA, USA, 30 October 2008*; pp. 9–16.
48. Roth, R.; Rambow, O.; Habash, N.; Diab, M.; Rudin, C. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of the ACL-08: HLT, Short Papers*; Association for Computational Linguistics: Valencia, Spain, 2008; pp. 117–120.
49. Seddah, D.; Chrupała, G.; Çetinoğlu, Ö.; Van Genabith, J.; Candito, M. Lemmatization and lexicalized statistical parsing of morphologically rich languages: The case of French. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*; Association for Computational Linguistics: Valencia, Spain, 2010; pp. 85–93.
50. Piskorski, J.; Sydow, M.; Kupś, A. Lemmatization of Polish person names. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, Prague, Czech Republic, 29 June 2007*; pp. 27–34.
51. Korenius, T.; Laurikkala, J.; Järvelin, K.; Juhola, M. Stemming and lemmatization in the clustering of finnish text documents. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, Washington, DC, USA, 8–13 November 2004*; pp. 625–633.
52. Kučera, K.; Stluka, M. Data processing and lemmatization in digitized 19th-century Czech texts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, Madrid, Spain, 19–20 May 2014*; pp. 193–196.
53. Eger, S.; Gleim, R.; Mehler, A. Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016*; pp. 1507–1513.
54. Lazarinis, F. Lemmatization and stopword elimination in Greek Web searching. In *Proceedings of the 2007 Euro American conference on Telematics and Information Systems, Faro, Portugal, 14–17 May 2007*; pp. 1–4.
55. Rakhimova, D.; Turganbayeva, A. Lemmatization of big data in the Kazakh language. In *Proceedings of the 5th International Conference on Engineering and MIS, Astana, Kazakhstan, 6–8 June 2019*; pp. 1–4.
56. Ozturkmenoglu, O.; Alpkocak, A. Comparison of different lemmatization approaches for information retrieval on Turkish text collection. In *Proceedings of the 2012 International Symposium on Innovations in Intelligent Systems and Applications, Trabzon, Turkey, 2–4 July 2012*; pp. 1–5.
57. Toporkov, O.; Agerri, R. On the Role of Morphological Information for Contextual Lemmatization. *arXiv* **2023**, arXiv:2302.00407.
58. Hafeez, R.; Anwar, M.W.; Jamal, M.H.; Fatima, T.; Espinosa, J.C.M.; López, L.A.D.; Thompson, E.B.; Ashraf, I. Contextual Urdu Lemmatization Using Recurrent Neural Network Models. *Mathematics* **2023**, *11*, 435. [[CrossRef](#)]
59. Gogoi, A.; Baruah, N. A Lemmatizer for Low-resource Languages: WSD and Its Role in the Assamese Language. *Trans. Asian-Low-Resour. Lang. Inf. Process.* **2022**, *21*, 1–22. [[CrossRef](#)]

60. Freihat, A.A.; Abbas, M.; Bella, G.; Giunchiglia, F. Towards an optimal solution to lemmatization in Arabic. *Procedia Comput. Sci.* **2018**, *142*, 132–140. [[CrossRef](#)]
61. Porter, M. The Porter Stemming Algorithm (1980). Available online: <http://tartarus.org/martin/PorterStemmer> (accessed on 9 September 2022).
62. Wikipedia Contributors. Gujarati Language—Wikipedia, the Free Encyclopedia, 2021. Available online: https://en.wikipedia.org/wiki/Gujarati_language (accessed on 4 December 2021).
63. Suba, K.; Jiandani, D.; Bhattacharyya, P. Hybrid inflectional stemmer and rule-based derivational stemmer for gujarati. In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP), Chiang Mai, Thailand, 8–13 November 2011; pp. 1–8.
64. Ameta, J.; Joshi, N.; Mathur, I. A lightweight stemmer for Gujarati. *arXiv* **2012**, arXiv:1210.5486.
65. Aswani, N.; Gaizauskas, R.J. Developing Morphological Analysers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages. In Proceedings of the LREC, Valletta, Malta, 17–23 May 2010.
66. Popat, P.P.K.; Bhattacharyya, P. Hybrid stemmer for gujarati. In Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010; p. 51.
67. Wallach, H.M.; Murray, I.; Salakhutdinov, R.; Mimno, D. Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 1105–1112.
68. Lau, J.H.; Newman, D.; Baldwin, T. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *EACL* **2014**, 530–539. [[CrossRef](#)]
69. Aletras, N.; Stevenson, M. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)—Long Papers*; Association for Computational Linguistics: Potsdam, Germany, 2013; pp. 13–22.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.