ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(27), DOI: 10.17485/ijst/2015/v8i27/81665, October 2015* ISSN (Online) : 0974-5645

Topic Modeling of E-News in Punjabi

Amandeep Verma* and Amandeep Kaur Gahier

Department of Computer Science and Engineering, Punjabi University Regional Campus for IT and Management, Mohali - 160062, Chandigarh, India; vaman71@gmail.com

Abstract

Topic Modeling refers to the act of discovering the theme of a document. Theme of a document provides an abstract view of the set of subjects (topics) addressed in the document. So, documents can be classified, arranged and searched according to their subjects using Topic Modeling. Topic Modeling has been the area of interest of most of the researchers from the fields of Text Mining, Natural Language Processing, and Machine Learning etc. Literature shows some techniques for generating theme out of a document. Most of the suggested Topic Models have been designed for English language. For Indian languages, particularly in Punjabi Language, such topic modeling is lacking in the literature. Although some Topic Summarization, Topic Tracking and Keyword Extraction systems has been developed for Punjabi Language, yet the technique of Topic Modeling is quite different from them. The paper presents a topic model for E-news in Punjabi Language. The idea of this topic model has been taken from the simplest and most basic probabilistic topic model; named LDA (Latent Dirichlet Allocation). This topic model finds the topics and their respective proportions present in the news text given as input to it. The theme generation process needs a Topic List Corpus at the backend of Topic Model. Such Corpus has been built containing Punjabi words commonly occurring in news articles, classified under different topic lists. The topic model has been tested on more than 1000 news articles for verification of its exactness. The values of various parameters attesting the quality of outputs given by topic model are quite satisfactory.

Keywords: Keyword Extraction, LDA, NLP, Probabilistic Topic Models, Topic List Corpus, Topic Lists, Topic Modeling, Topic Summarization, Topic Tracking

1. Introduction

Evolution of the Internet has provided us with a wealth of information. One can find lot of documents related to a subject, but in limited span of time, human mind is not able to search all of them so as to get required information. So, rather than finding documents through keyword search alone, it is better to find the theme of the document and then search on documents of the related theme only. It will reduce the number of documents, as it filters the document on the basis of theme, and thus search will be more effective. A topic model[1–7] attempts to discover the theme of a document under consideration, the discovered theme of a document can give an idea if that document has the content relevant to the problem. Apart from this, topic models can classify and organize various documents according to their themes.

In the literature[1–7], most of the topic models have been designed for English language. Considering their significance, as reported in the literature, a topic model for E-news in Punjabi language have been proposed, because very less amount of work has been done in the field of topic modeling for Indian languages (especially Punjabi language). The idea of its design has been taken from the basic probabilistic model, LDA (Latent Dirichlet Allocation)[8–11]. LDA contains all the essential steps involved in the task of topic modeling, so it serves as a reference framework for all other domain specific topic models to be developed.

To develop a topic model, a source (topic lists) containing words occurring in a language is needed, using which theme of a document can be generated. The scope of Punjabi language is so vast, as it contains words from simple Punjabi literature to complex words of Gurbani.

* *Author for correspondence*

Punjabi News is the source which covers a wide scope of words occurring in Punjabi vocabulary. So, the proposed topic model has been designed to extract the theme of Electronic News texts. The outputs of the proposed model were evaluated on parameters precision, recall, lack and sufficiency[12].

1.1 Review of Literature

The topic models developed in recent years have been classified into three categories[1]:

- Set theoretical methods (these methods represent the documents as set of words and use set theoretic operations to do further processing).
- Algebraic models (These methods represent documents and queries using vectors/matrices/tuples).
- Probabilistic models (These models treat the process of topic modeling as a problem of probability. Theorems of probability are often used in these methods).

Probabilistic models have been the area of interest for most of the research in the last decade, as they are considered better than the other approaches of using sets and matrices for topic modeling. The idea of the proposed topic model for the Punjabi language has also been taken from a probabilistic model named, Latent Dirichlet Allocation (LDA). LDA is the simplest and the most popular probabilistic model[2,3]. As it comprises of all the basic steps involved in the extraction of the theme of a document, it serves as a basic reference framework for other probabilistic topic models.

Most of the probabilistic topic models extract the theme from an article in two steps[2,8,9]: Generative process and Statistical inference. Generative process explains how the text would have been generated and statistical inference infers the results of generative process using statistical methods.

The tools related to Topic Modeling are also reviewed so as to gain knowledge about implementing a Topic Model. Machine Learning for Language Toolkit (MALLET)[13,14] is a Java based package, which when supplied a topic list corpus (containing words of certain area e.g. infections), it creates a new topic model. The topic model hence created is able to generate theme of the English texts. Stanford Topic Modeling Tool (TMT)[19,20] is also a Java Based Package. This is a topic model developed at Stanford University and it generated theme for English Language.

2. The Proposed Model

The Topic Model for Punjabi Language aims to extract the theme of given Punjabi text. The idea of the strategy being followed for topic modeling has been taken from LDA[9]. If a collection of words of Punjabi literature is available and these words are arranged into different lists in such a way that the words with similar meaning and falling into same category are placed in same list, then each of the word from the input text can be assigned a category to which they belong (per word topic assignment) leaving helping verbs etc. This process is called Generative Process, as it justifies how the input text would have been generated. After this process, the proportion of involvement of each topic in the input text is computed. This second process is termed as Statistical Inference Process. A histogram describing the topic proportions present in the document can be formed. The histogram thus created can tell the collection of topics that are involved in the text (example 10% about education, 30% about health, and so on). This information can help to a great extent in finding out what theme the text contains. The idea of the proposed process is illustrated in Figure 1. Here, a simple Punjabi story is considered and the topic collection contains topics "ਹਾਦਸਾ, ਮਣਿਤੀ, ਰੱਬ, ਕੰਮ, ਪੰਛੀ" and the words that belong to these topics. Each of the word from the Punjabi story (considered as input) is matched against all words in each topic list available.

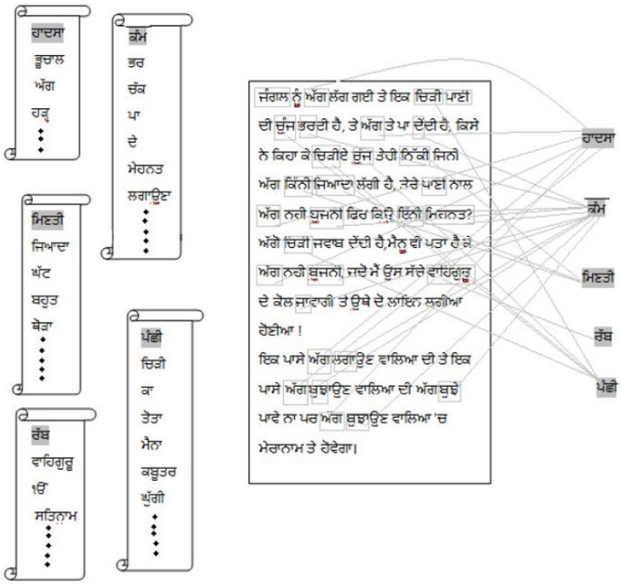

Figure 1. Topic modeling in Punjabi language.

Words falling in one category, say "ਹਾਦਸਾ" are counted, similarly for other categories and a histogram containing all the matched topics and the proportions of their presence can be formed.

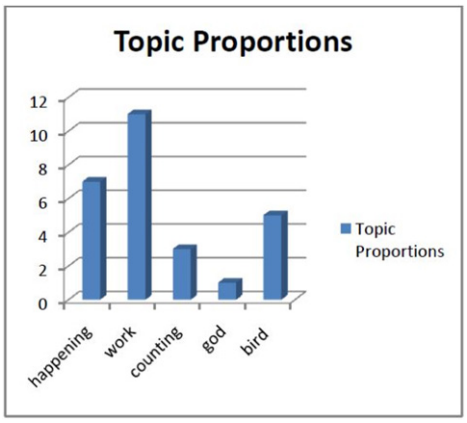

Figure 2. Desired form of output of proposed model.

The desired output histogram from the topic model is represented in Figure 2. It shows the list of topics on the horizontal axis that occur in the Punjabi text given input to the topic model. The vertical axis shows the proportion of each of the identified topic present in the text. The availability of this information clearly shows the theme of the text.

3. Design

To accomplish the proposed model, the main requirement is to design a topic list that contains collection of various topics that occur frequently in news articles and words corresponding to those topics. These topic lists can be prepared manually or some learning algorithm can be employed to generate such collection for the required model.

The topic model for Punjabi language proposed in the study needs to perform following two major tasks, in order to find the theme of the text given as input to it:

Step 1: Matching the words of the Punjabi news text to the topic list corpus.

Step 2: For each topic list, counting the words matching to that particular topic list.

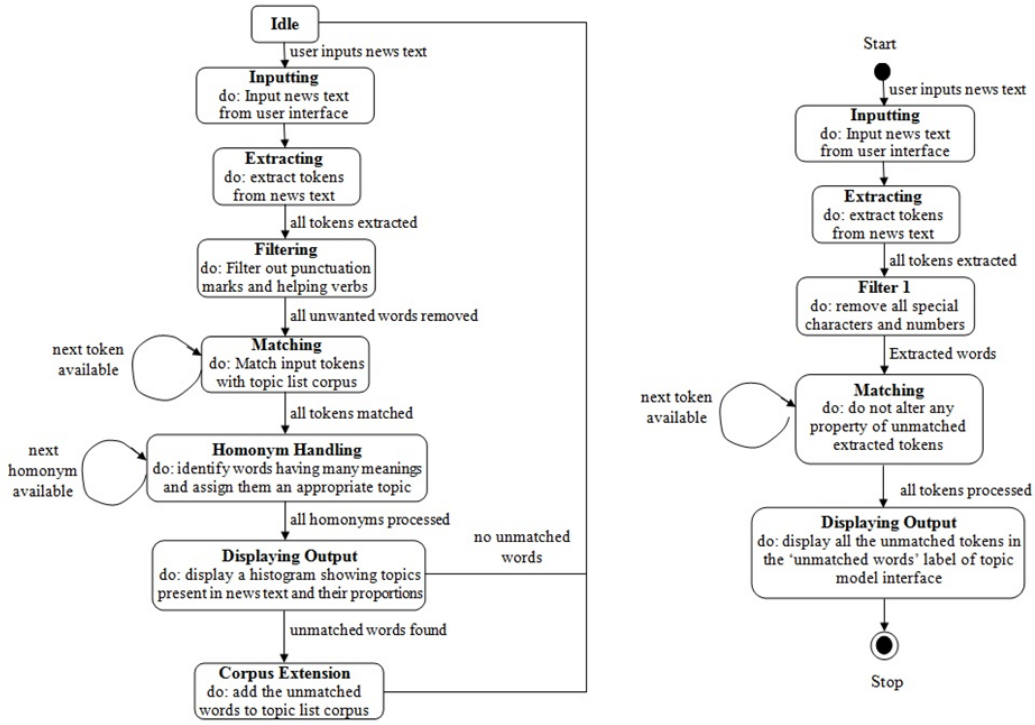

Figure 3. State chart diagram for scenario 1, 2.

Table 1. Type of treatment to different tokens in matching process

Type of token	Treatment given to token
The token is not a Punjabi word. It is composed of numbers or special characters.	Token is discarded
The token is a helping verb in Punjabi Language.	Token is discarded
The token is not present in any list of corpus.	Token is not assigned any topic
The token is matched to one list of corpus.	Token is assigned exactly one topic
The token is matched to more than one list of corpus (i.e. the token is a homonym)	Token is assigned more than one topics (this needs additional steps to resolve the ambiguity among assigned topics)

The *Step 1* matches input Punjabi tokens to each and every word present in the whole topic list corpus. In this step, each of the token occurring in input can be treated as one of the cases mentioned in Table 1.

The aim of this step is to find out the topics which originated this text, due to this reason *Step 1* can be named as Generative Process.

The *Step 2* aims to count the frequencies of each topic that is found present in the input text. If any homonyms (words matched to more than one topic list simultaneously) are present in the text, they are modeled to one appropriate topic list before the counting of frequencies. *Step 2* infers the share of a topic in the theme of the text, so it can also be called Inference Process.

In Punjabi news text, a lot of unnecessary names of people, objects, helping verbs, special characters and numeric details are present which are not required to be processed for the theme generation process. Such tokens are eliminated from the input text using filters. The Topic Model being proposed in this study is a complex system. It

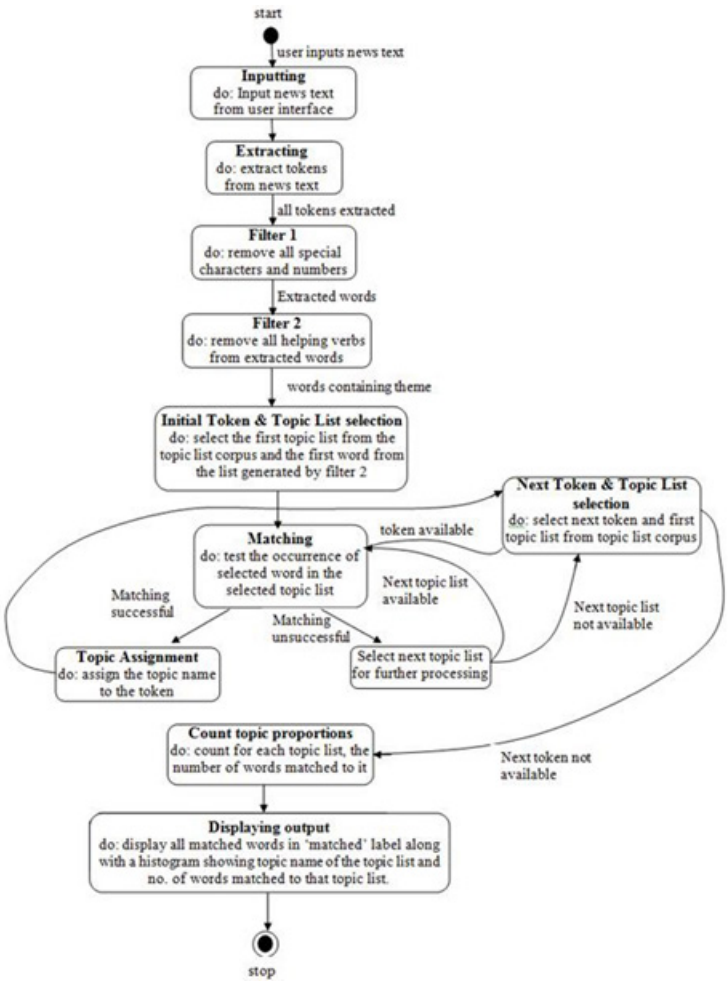

Figure 4. State chart diagram for scenario.

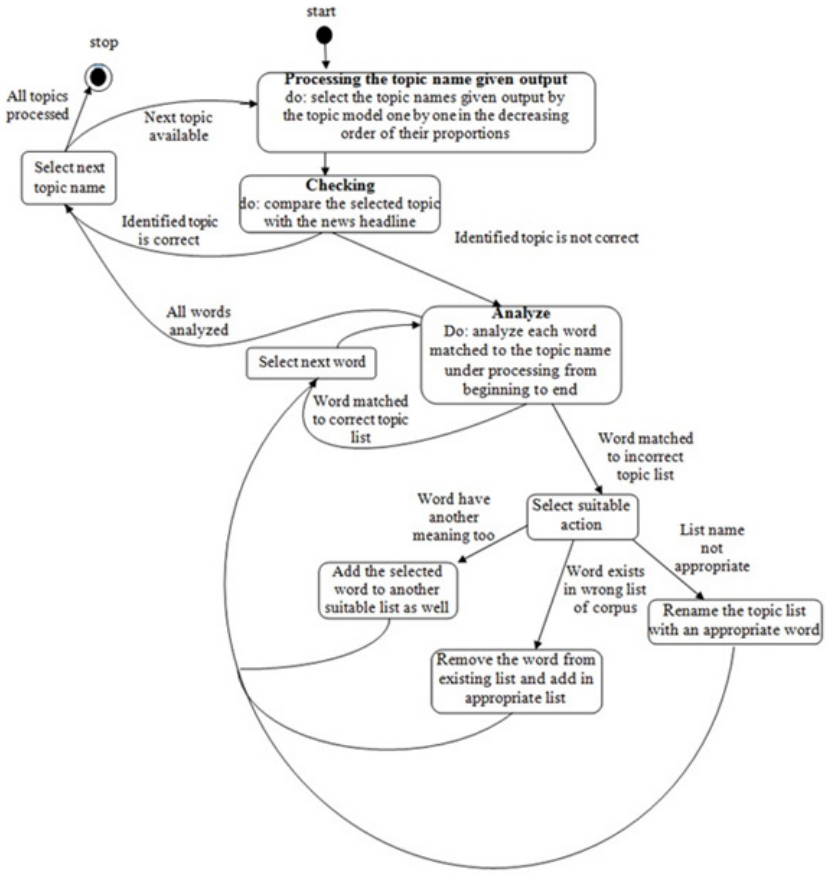

Figure 5. State chart diagram for scenario 7.

involves a lot of operations to be performed repeatedly on the data given as input to it. The complete topic model can be explained in 8 possible scenarios given in the Table 2.

Table 2. Different scenarios of topic model

Scenario 1	Mainline sequence
Scenario 2	Input text is not appropriate (not in Punjabi language/ consist only special characters)
Scenario 3	Input news text contains words, such that each and every word exists in topic list corpus.
Scenario 4	Input news text contains words, such that any word does not exist in topic list corpus.
Scenario 5	Input news text contains words having more than one meaning (homonyms)
Scenario 6	Input news text contains a mixture of words existing in corpus, not existing in corpus and homonyms
Scenario 7	Wrong input received, make corrections in topic lists
Scenario 8	Extend topic lists by adding new words in them and add more topic lists in corpus

All of these scenarios are important for understanding the operation of the system, but in the present paper some of them are elucidated.

Mainline sequence means the normal operation of the topic model, when no exception is occurred. The topic model is supposed to input the news text first, and then extract the tokens from the input. This extraction is followed by filtration process (removal of all unnecessary words which do not contribute to the theme). Then the matching process begins, which is the most complex and important part. After words are matched, some words matched to more than one list are considered as homonym and they are given special treatment to resolve the ambiguity among topics. After that, system enters into a state when output is displayed and the user can extend the corpus using unmatched words left. Figure 3 shows State Chart diagrams of Scenario 1 and 2.

Scenario 5 explains the case when input news text contains words having more than one meaning (homonyms). Its state chart diagram is given in Figure 4.

Scenario 7 explains the case when wrong output is received and corrections are made in topic lists. Figure 5 explains the state chart diagram for this scenario.

For the data transformation details, the Topic Model is explained using Data Flow Diagram as shown in Figure 6.

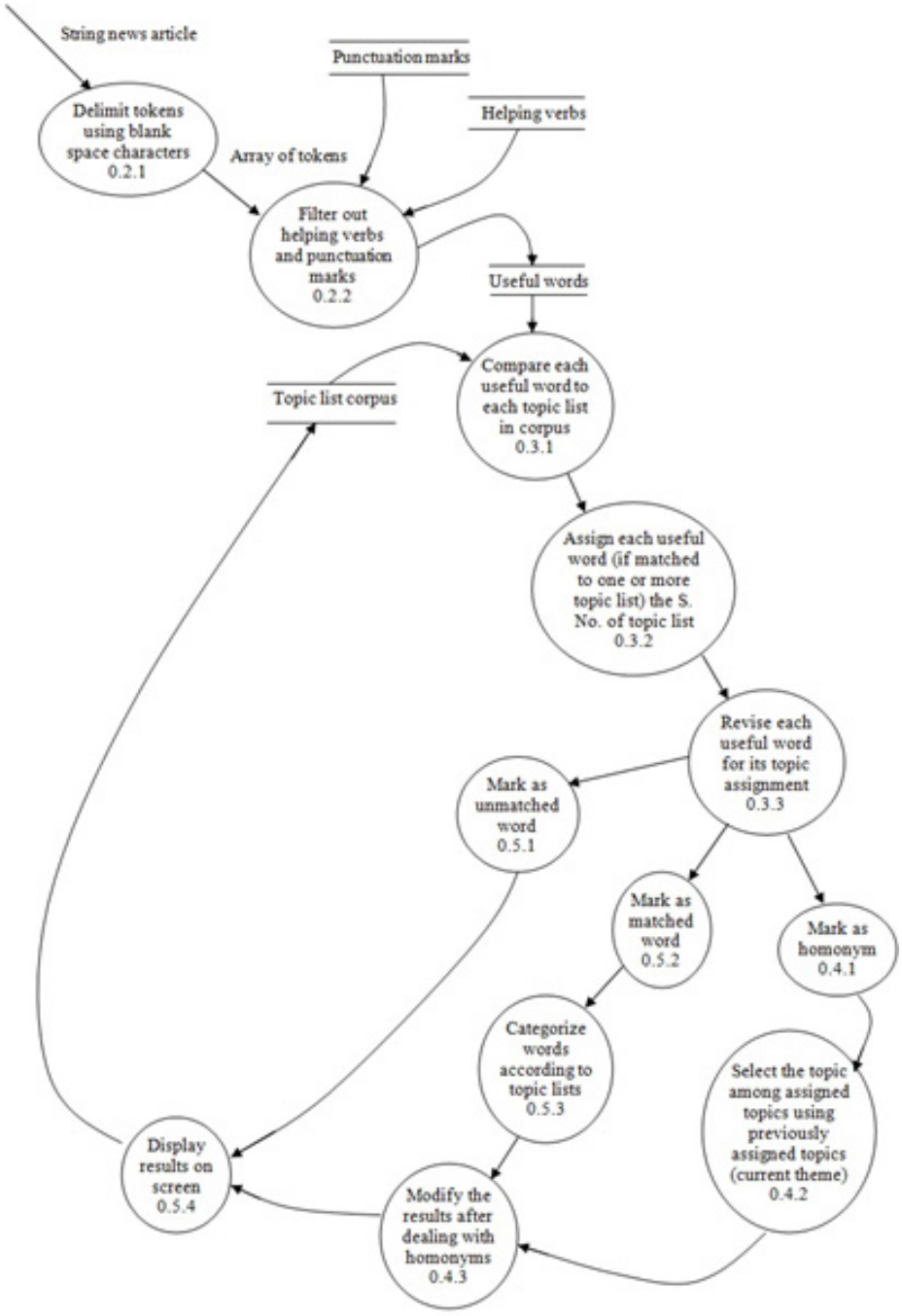

Figure 6. Data flow diagram of topic model.

4. Implementation

The topic model has been implemented using ASP.NET and the scripting part is done using C#. The Topic List Corpus is built using SQL Server Database and is connected to the Topic Model. The operation of Topic Modeling has been carried out by different methods of three C# classes namely "InputNewText", "TopicModeling" and "OutputNewsText". A sample output of the topic model can be shown using the Figure 7.

The Topic List Corpus used for this implementation is a collection of limited words only. The Corpus contains 48 topic lists containing 1017 Punjabi words in them.

5. Results and Discussions

The topic model has been tested on more than 1000 news texts. The testing is done by comparing the obtained output

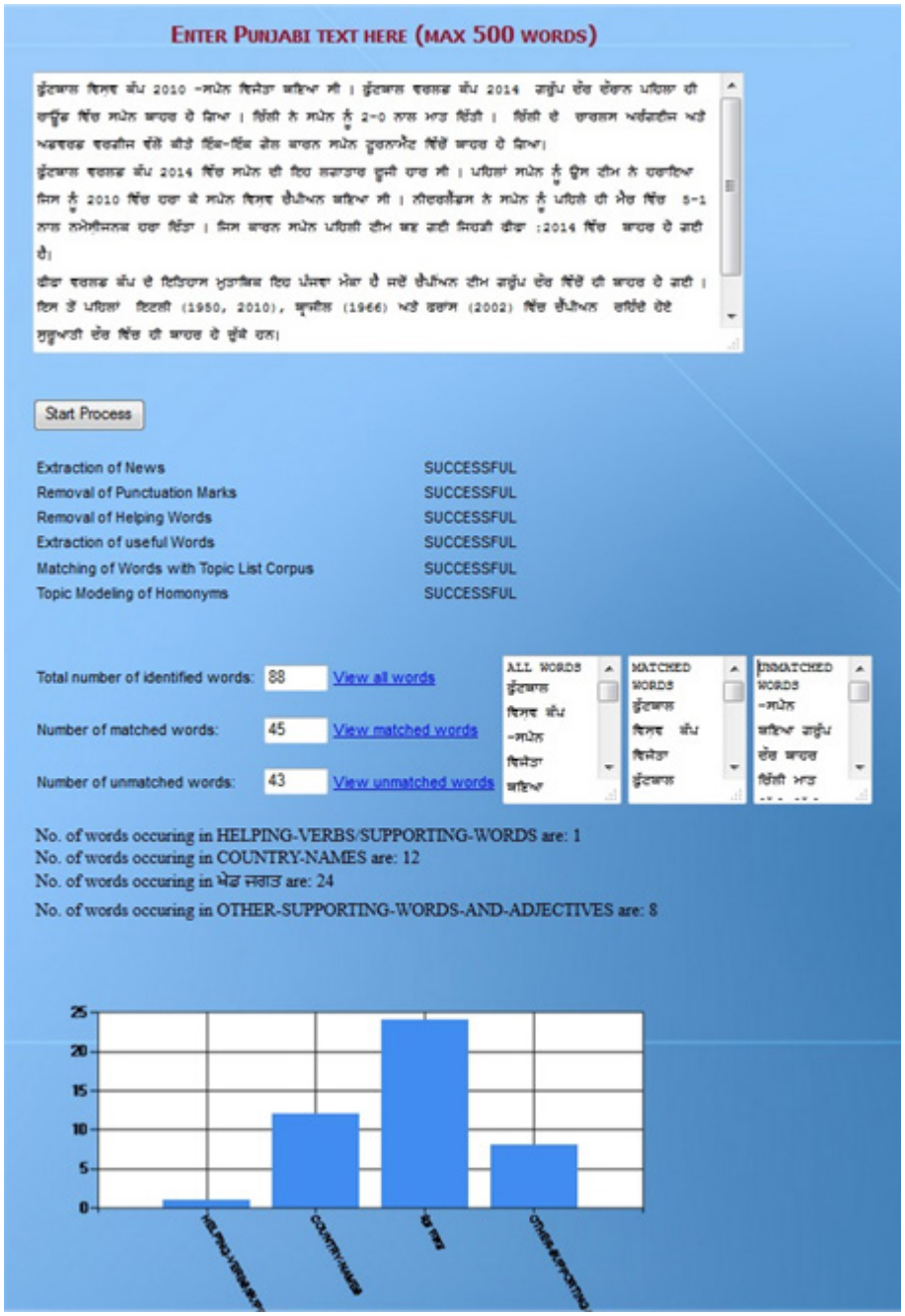

Figure 7. A sample output.

theme from the topic model with the corresponding news headline as well as the content of the news text. The performance measures used in the study are:

Precision: Precision measures the ratio of relevant output instances to the total instances obtained from the output[12].

$$Precision\ P1 = \frac{News\ Headline\ topics\ correctly\ identified\ by\ Topic\ Model}{All\ topics\ in\ news\ headline} \text{x}100$$

$$Precision\ P2 = \frac{Correctly\ identified\ topics\ of\ Output}{all\ topics\ given\ output\ by\ topic\ model} \text{x}100$$

$$Precision = \frac{Precision\ P1 + Precision\ P2}{2}$$

Recall: Recall is the ratio of relevant output instances to the total instances[12].

$$Recall = \frac{no.of\ correct\ topics\ obtained\ by\ the\ system}{total\ number\ of\ topics\ that\ have\ been\ used\ for\ testing}$$

F-measure: The parameter F- measure gives the overall performance of the topic model[12].

$$F\text{-}\ measure = \frac{2PR}{R+P}$$

where P and R stands for Precision and Recall respectively.

Sufficiency: This parameter shows the amount of content from input news text that has been successfully matched to the topic list corpus.

$$sufficiency\% = \frac{no.\ of\ matched\ words}{total\ no.of\ words} x100$$

Lack: Lack of words in topic list represents the need to include more words from Punjabi vocabulary so that all of the input words can be matched to some topic list of topic list corpus.

$$lack\% = \frac{no.\ of\ unmatched\ words}{total\ no.\ of\ words} x100$$

The topic model has been tested with 1000 news texts, out of which a set of 150 news texts come under training data set and rest 850 come under testing data set. The training data set has been used to populate the Topic List Corpus with different Punjabi words.

Figure 8 shows when the test results of the topic model are verified according to the headline of the news.

From the Figure 8, it can be seen that the curve of test cases giving correct results is increasing after training phase.

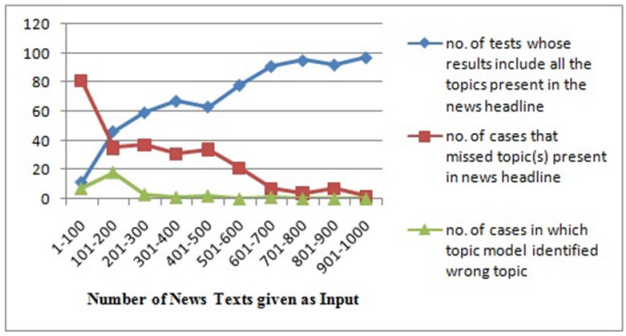

Figure 8. Verification of results according to headline.

A graph in the Figure 9 shows the test results when the outputs are verified according to content of news text. It is clear that the topic model results a theme which is more detailed than the news headline. Moreover, number of cases resulting accurate themes increase monotonically (except at one place, such irregularities occur when the topic model is tested with such articles, whose words are not included in topic list corpus).

The results of the topic model can also be analyzed according to the type of news texts. Figure 10 shows the effect of number of tests done on Precision for different types of news texts.

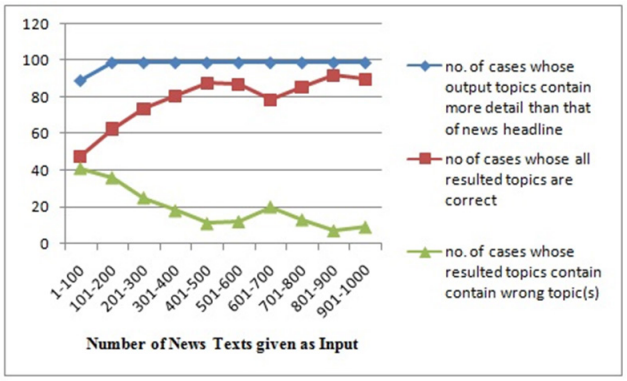

Figure 9. Effect of training on recall and positive tests.

It has been also observed that more are the number of tests done with a particular type of news, more is the recall value and more are the number of matched words during the process of topic modeling. This can be seen in the Figure 11 given below.

The average Sufficiency% of the topic list corpus has found to be 39.65%. This means average 39.65% words from the news text given input to the Topic Model will be matched to the Topic List Corpus and these words will participate in theme generation process. Similarly, Lack% of the topic model is found to be 60.35% which means, 60.35% words remain unmatched during Topic Modeling.

The aim behind keeping sufficiency% up to 40% is that, any language contains a number of words (supporting or helping verbs, names of people and places etc.) which doesn't affect the theme of a text. If all of such supporting words and names were included in the topic lists, then the Topic model will start giving an output which would be more similar to the translation of the input rather than the theme of the same.

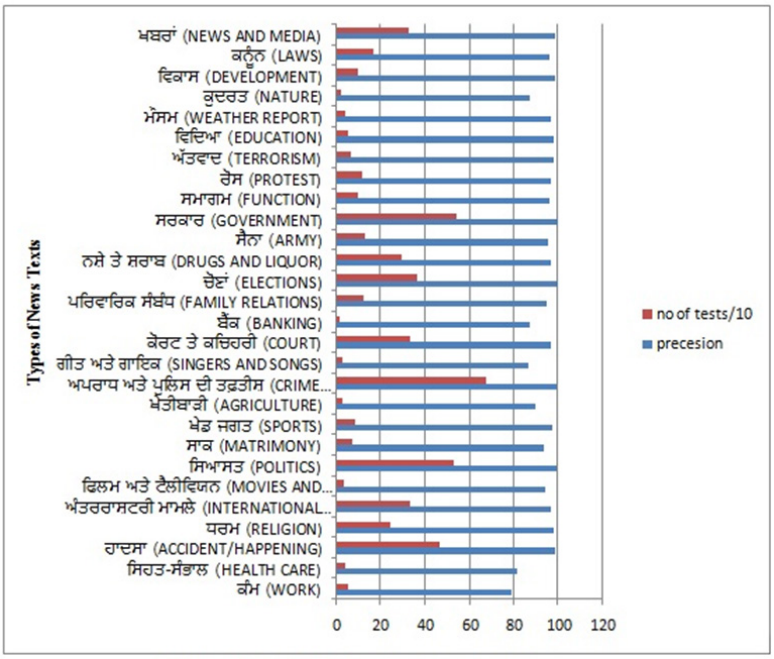

Figure 10. Effect of number of tests done on precision.

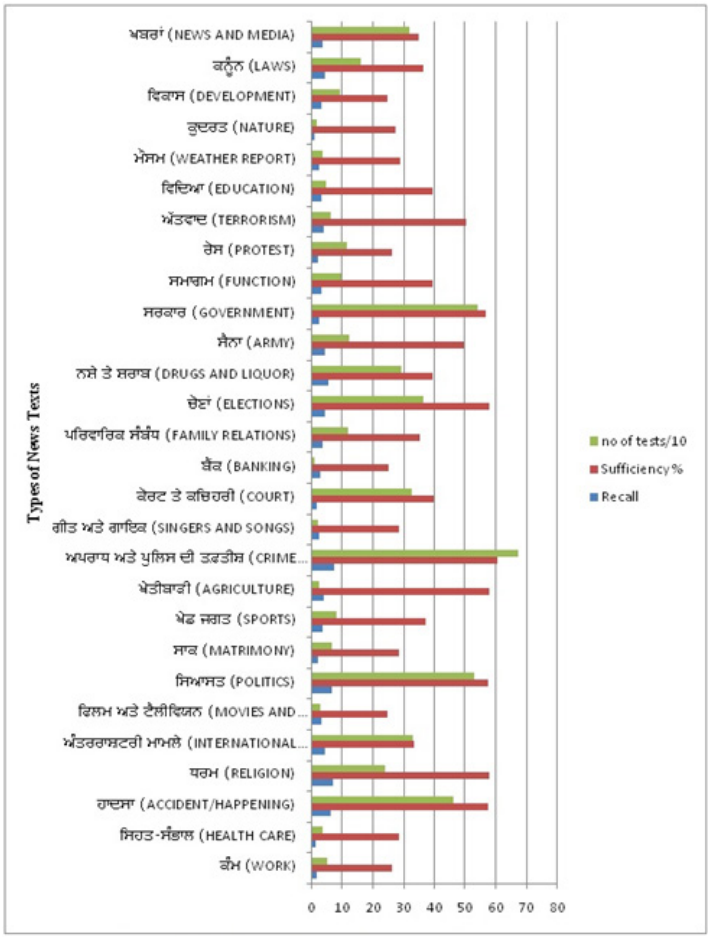

Figure 11. Effect of number of tests done recall and sufficiency.

6. References

1. Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. Probabilistic author-topic models for information discovery. Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004. p. 306–15.
2. Blei D. Probabilistic topic models. Communications of the ACM. 2012; 55(4):77.
3. Probabilistic topic models. Communications of the ACM. 2015; 55(4):77–84. Available from: http://dl.acm.org/citation.cfm?doid=2133806.2133826
4. Barbieri N, Manco G, Ritacco E, Carnuccio M, Bevacqua A. Probabilistic topic models for sequence data. Mach Learn. 2013; 93(1):5–29.
5. Blei D, Carin L, Dunson D. Probabilistic topic models. IEEE Signal Process Mag. 2010.
6. Graham S, Milligan I. Topic modeling with the Stanford TMT. Available from: http://themacroscope.org
7. Ramage D, Rosen E. Stanford topic modeling Toolbox.The Stanford Natural language Processing Group. Available from: http://nlp.stanford.edu/software/tmt/tmt-0.2/
8. Steyvers M, Griffiths T. Handbook of latent semantic analysis: A road to meaning. Laurence Erlbaum. 2007.
9. Blei D, Ng A, Jordan M. Latent dirichelet allocation. The Journal of Machine Learning and Research. 2003; 3:993–1022.
10. Chen E. Introduction to latent dirichlet allocation - Edwin Chen's Blog. Available from: http://blog.echen.me/2011/06/27/topic-modeling-the-sarah-palin-emails
11. Reed C. Latent dirichlet allocation: Towards a Deeper Understanding. 2012. p. 1–13.
12. Jizba R. Measuring search effectiveness. 1st ed. 2008. Available from: https://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching_-_Recall_Precision.pdf
13. Mimno D. Machine learning with MALLET. 1st ed. Available from: http://mallet.cs.umass.edu/mallet-tutorial.pdf
14. McCallum A. MALLET: A machine learning for language toolkit. 2002. Available from: http://mallet.cs.umass.edu