Topic Modeling on Twitter Data

*A project report is submitted to the department of*

***Information Technology***

*of*

***International Institute of Information Technology Bhubaneswar***

*in partial fulﬁlment of the requirements*

*for the degree of*

***Bachelor of Technology***

*by*

***Sarita Sahu***

*(ID- B416042)*

*under the supervision of*

***Prof. Rakesh Kumar Lenka***

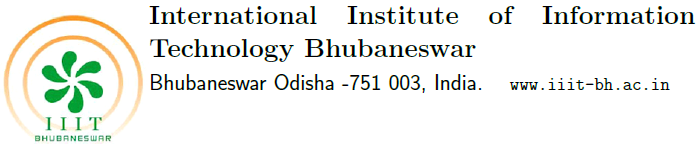


**Information Technology**

**International Institute of Information Technology Bhubaneswar**

**Bhubaneswar, Odisha - 751003, India**

**2020**

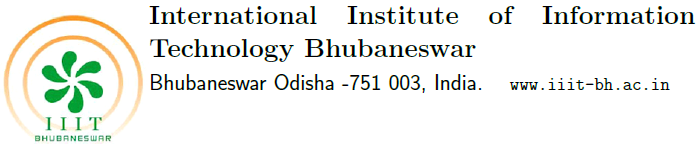


25th May, 2020

**UNDERTAKING**

I declare that the work presented in the report titled Topic Modeling on Twitter Data, submitted to the Department of Information Technology, International Institute of Information Technology, Bhubaneswar, for the award of the Bachelors of Technology degree in the Information Technology, is my original work. I have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, I accept that my degree may be unconditionally withdrawn.

***Sarita Sahu, B416042***



25th May, 2020

**CERTIFICATE**

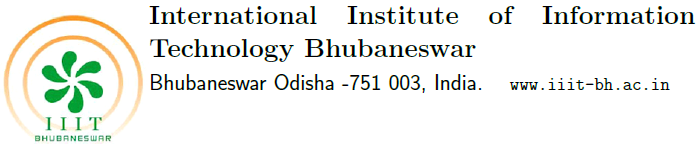
This is certify that the work in the project titled Topic Modeling on Twitter Data by Sarita Sahu is a record of an original work carried out by her under my supervision guidance in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology. Neither this project nor any part of it has been submitted for any degree or academic award elsewhere.

**Rakesh Kumar Lenka**

**CERTIFICATE OF INTERNSHIP**

****

|  |
| --- |
|  |



25th May, 2020

**ACKNOWLEDGMENT**

The elation and gratification of this seminar will be incomplete without mentioning all the people who helped me to make it possible, whose gratitude and encourage were invaluable to me. I would like to thank God, almighty, our supreme guide, for bestowing his blessings upon me in my entire endeavor. I express my sincere gratitude to Prof. Swati Vipsita and Rakesh Kumar Lenka for their guidance and support and students of my batch for their suggestions and support.

***Sarita Sahu***

**CONTENT**

Page No.

1. Abstract 1
2. Chapter 1 Introduction

* Twitter 2
* Covid19 2

1. Chapter 2 Overview 3
2. Chapter 3 Requirements

* Dataset 4
* Programming Language 5
* Libraries 5
* Model 5

1. Chapter 4 Implementation

* Data Preprocessing 7
* Implementation 8
* Modifications 10

1. Chapter 5 Improvising 12
2. Chapter 6 Conclusion 13
3. Bibliography 14

**ABSTRACT**

Topic modeling is an unsupervised machine learning technique that is capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. It helps in discovering hidden topical patterns that are present across the set of documents. A document can be a single sentence or a group of sentences or instances. In this project, we have used tweets from Twitter on recent situation in India to perform topic modeling to see on what topic people are mostly talking about.

**Chapter 1**

**INTRODUCTION**

* 1. **Twitter**

Twitter is social networking service founded in the year 2006 in California. It is the place where users post and interact with other users through messages known as “tweets”. Hence it’s a gold mine of huge amount of data. It is very much specific. Any person in this world can create an account here and connect with people around the world. Almost every user’s tweets are completely public and pull able. Twitter provides its API to users to access tweets through complex queries.

* 1. **COVID 19 (Novel Corona virus)**

COVID-19 is a disease caused by a new strain of corona virus. It originated from Wuhan, China and has spread all over the world rapidly at an alarming rate. People all over the world have been talking about this, sharing their thoughts, wishes, helping others in this crucial time.

So, for this project I have pulled tweets from the Twitter related to this issue to analyze and get some insights from it.

**Chapter 2**

**OVERVIEW**

The objective of the project is to understand the topics on which the people of a certain region are discussing over a predefined period of time. Here, an algorithm named Latent Dirichlet Allocation (LDA) is applied on the dataset to extract the hidden topics or texts in the corpus or documents (tweets). It is the most common topic model algorithm.

Latent means hidden, Dirichlet denotes Dirichlet distribution or multivariate beta distribution that belongs to the family of continuous multivariate probability distributions parameterized by a vector of α of positive reals and Allocation refers to the allocation of topics to each of the documents.

**Chapter 3**

**REQUIREMENTS**

**3.1 Dataset**

Since the topic is about corona virus, we have pulled data from Twitter related to the current situation through a specific query. Twitter provides APIs to developers to help them access data in various ways. Using the standard stream API, we can extract data from the past 7-8 days.

Twitter provides access and consumer secret keys to have a secure connection for pulling data. It is possible to pull tweets specific to a geo location, user account, number of tweets per page, types of tweets, retweets, language of tweets and many more features.

I have successfully able to pull tweets of users from the Delhi, India location using the longitude and latitude of Delhi with a radius of 50km. The data is in the English language and no retweets have been pulled. It ranges from 2nd of May till 21st of May, 2020. It contains nearly 2800 tweets, a mixture of popular and unpopular tweets.

The tweets are specific to a list of hash tags related to covid19 situation in the world. This list contains twenty hash tags which are used to pull tweets from Twitter.

**3.2 Programming Language**

To perform topic modeling, the first step is to obtain data. Next, the data needs to be cleaned for which Python or any other language can be used which provides libraries to perform the required tasks.

Python provides various libraries which makes it possible to perform machine learning and other related tasks efficiently.

**3.3 Libraries**

The following libraries from Python are necessary to perform topic modeling on tweets:

* Pandas
* Tweepy
* Nltk
* Re
* Os
* Glob
* Sklearn
* matplotlib

**3.4 The Model**

LDA is based on probabilistic graphical modeling. The document is thought to be a distributed over topics and each topic contains a set of words that closely relate to the document.

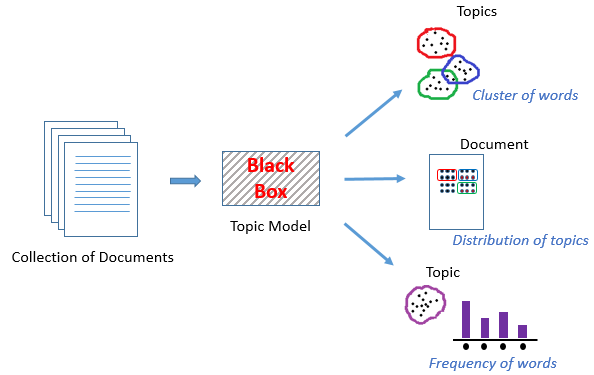


Fig 1.Topic Modeling Structure

The above figure shows how the documents are classified into various topics. In this algorithm, the model is built on the bag of words matrix as input and we have to decide the number of topics to be generated and how many top words are to be assigned to each topic.

Hence it produces two matrices: one, topic-word matrix and two, document-topic matrix. Topic-word matrix contains the words in each topic with their associated weights and document-topic matrix contains the probability distribution of topics over each document.

**Chapter 4**

**IMPLEMENTATION**

**4.1 Data Preprocessing**

The data is a mixture of many components. It contains a mixture of special characters, non-ASCII characters, emotion icons, URLs, mentions, etc. Before applying the model on the data, it needs to be cleaned very carefully in order to get proper results from model. If we fit the raw data into the model, the result obtained would be meaningless.

So the tweets are cleaned by removing “stopwords” (the words which are very common and occur in every sentence like the,an,a,this ,etc),,removing the mentions(like @Modi),removing URLs,hashtags(like #stayhome) and other characters and lemmatization of words (which means generalizing the words like gets -> get, getting ->get ,etc). We also remove those words which appear very frequently or less frequently. After this, if the tweet contains less than 10 characters, they are removed.

Now, the data is ready to be fit into the model.

**4.2 Implementation**

The tweets are cleaned and are fit to the model with number of topics as 10 and number of top words as 10.The other parameters are set to default.

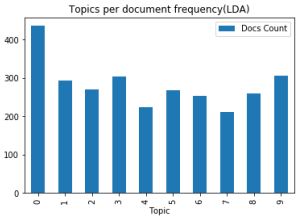


Fig 2 Frequency chart

The above figure shows the frequency of the distribution of topics over the documents. It is observed that topic 0 has highest number of documents and the others have nearly the same.

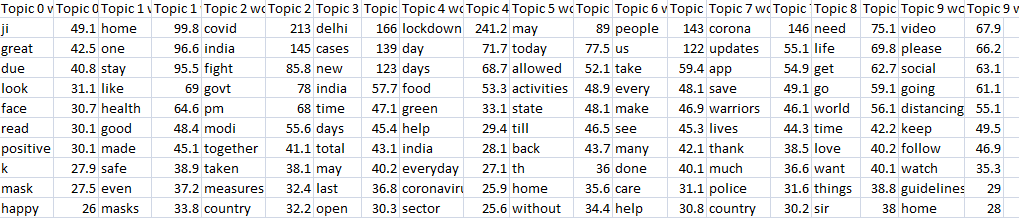


Fig 3 Topic-word matrix

The figure above shows the distribution of words over topics with their weights. It is observed that topic 1 tell us about staying home healthy, topic 9 about follow social distancing, topic 4 about days and zones,etc.

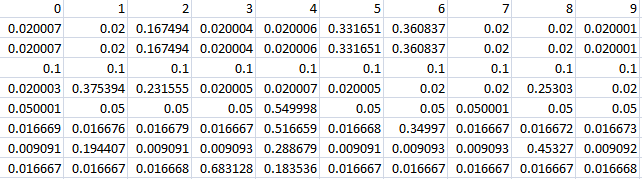


Fig 4. Document-topic matrix

A glimpse at the document-topic matrix shows how much the topics are related to a document and the topic with the highest probability represents the topic of the document.

But it is noticed that some documents have the same probability for all the topics which means that document is not properly fit to the model.

**4.3 Modifications**

As the above model was not able to classify the documents properly there needs to be perform some modifications. Hence, the documents that show same probability of topics are removed from the dataset. And the count of such documents was found to be 263 which mostly belonged to topic 0. Rest of the parameters was kept constant as before and the model was run again.

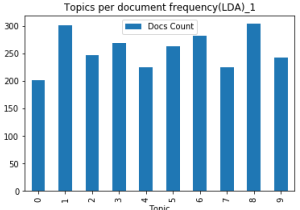


Fig 5. Frequency chart

The frequency of documents changed for each topic and some they remained almost same.

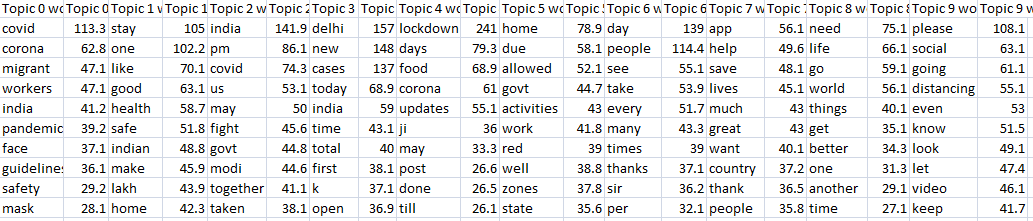


Fig 6.Topic-word matrix

Now the words have changed for the topics and look a little more meaningful. The topic 0 tells about migrant workers and guidelines, topic 2 about staying safe at home and so on.

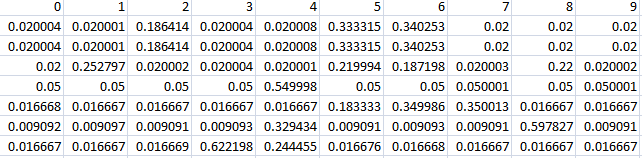


Fig 7 Document-topic matrix

And the document-topic matrix is shown above. For the model to be good there should be only one prominent topic that represents a document. Therefore, only topic ought to carry the highest probability in the document.

**Chapter 5**

**IMPROVISING**

The model needs to be further improved to get better results. Though I have tried modifying a lot, still there are some flaws which I couldn’t be remove. Or maybe the data is not so good that the model couldn’t perform well. The tweets may have been very vague or may contain anomalies. It was also found that many of the tweets contained the main topic as in form of a video or image, so the textual part of the tweets were very less or negligent.

In such cases, the tweets need to be more meaningfully pulled so as to get some valuable information from them. Moreover, the location we chose was small compared to a country hence such vagueness. One may try using some other modeling algorithms and see if that can help us getting better results.

**Chapter 6**

**CONCLUSION**

It was nice experience working on this subject/field. This was the first time I tried working on a project all by myself with the help of teachers and friends and it was very good. My brother has also helped me a lot in this project. I had to watch videos and study a few articles to get to know what this topic was about and found it to be very interesting.

The best part was the implementation as it was challenging for me to decide on the result obtained whether it was to be considered right or wrong or needs modification or improvement. Analyzing data and the results are fun and help a lot in the future.

**BIBLIOGRAPHY**

1. <https://scikitlearn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html>
2. <https://medium.com/mlreview/topic-modeling-with-scikit-learne80d33668730>
3. <https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf>
4. <https://towardsdatascience.com/topic-modeling-quora-questions-with-lda-nmf-aff8dce5e1dd>