TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

THESIS NO.: 076MSCSK020

Sentiment Analysis and Topic Modeling on News Headlines

by

Vijay Yadav

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SYSTEM AND KNOWLEDGE ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

LALITPUR, NEPAL

September, 2022

Sentiment Analysis and Topic Modeling on News Headlines

by

Vijay Yadav

076MSCSK020

Thesis Supervisor

Prof. Dr. Subarna Shakya

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer System and Knowledge Engineering

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk Campus

Tribhuvan University

September, 2022

COPYRIGHT ©

RECOMMENDATION

The undersigned certify that they have read and recommended to the Department of Electronics and Computer Engineering for acceptance, a thesis entitled **"Sentiment analysis and topic modeling on news headlines"** submitted by **Vijay Yadav** in partial fulfillment of the requirement for award of the degree of "**Master of Science in Computer System and Knowledge Engineering"**.

……

Supervisor: Prof. Dr. Subarna Shakya,
Department of Electronics and Computer Engineering,
Institute of Engineering, Pulchowk Campus

……

External Examiner: Om Bikram Thapa,
Chief Technology Officer,
Vianet Communications Pvt. Ltd.

……

Committee Chairperson: Assoc. Prof. Dr. Nanda Bikram Adhikari,
Department of Electronics and Computer Engineering,
Institute of Engineering, Pulchowk Campus

Date: September, 2022

DECLARATION

I declare that the work hereby submitted for Masters of Science in Computer System and Knowledge Engineering at IOE, Pulchowk Campus entitled "**Sentiment Analysis and Topic Modeling on News Headlines**" is my own work and has not been previously submitted by me at any university for any academic award. I authorize IOE, Pulchowk Campus to lend this thesis to other institution or individuals for the purpose of scholarly research.

Vijay Yadav
076-MSCSK-020
Date: September, 2022

DEPARTMENTAL ACCEPTANCE

The thesis entitled **"Sentiment Analysis and Topic Modeling on News Headlines",** submitted by **Vijay Yadav** in partial fulfillment of the requirement for the award of the degree of "**Master of Science in Computer System and Knowledge Engineering**" has been accepted as a bonafide record of work independently carried out by him in the department.

Prof. Dr. Ram Krishna Maharjan
Head of the Department
Department of Electronics and Computer Engineering,
Pulchowk Campus,
Institute of Engineering,
Tribhuvan University, Nepal

ACKNOWLEDGEMENT

I am grateful to my thesis supervisor Prof. Dr. Subarna Shakya for providing his invaluable feedback and guidance throughout this thesis work.

I would like to express my gratitude to our coordinator Dr. Nanda Bikram Adhikari for providing a conducive environment and assisting with necessary resources to carry out the thesis work.

I would also like to thank Department of Electronics and Computer Engineering, Pulchowk Campus, for providing me an opportunity to do the thesis and assisting with all necessary support and guidance.

ABSTRACT

In today's world of competitiveness, sentiment analysis has wide range of applications from medical perspective to examine the mental situation of the person to entertainment industry, corporates, politics and so on in order to examine the perspective and views of the people towards their product. News media play vital role in shaping the views of public regarding any product or people. The dataset used for this thesis is headlines dataset of one of the leading new portals of India i.e., Times of India. Both supervised and unsupervised techniques would be used to perform the analysis on the dataset. The thesis has two aspects i.e., first, sentiment analysis for which supervised technique Bi-LSTM will be used and second, topic modeling for which unsupervised techniques LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Analysis) will be compared, and then the best performing algorithm will be used for topic classification. The topics identified will be used to classify the dataset so that prediction of topic for particular headline can be done.

Keywords: sentiment analysis, topic modeling, data visualization, News headline, Bi-LSTM, LDA, LSA

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

LSTM	Long Short-Term Memory
Bi-LSTM	Bidirectional Long Short-Term Memory
RNN	Recurrent Neural Network
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
ML	Machine Learning
NLP	Natural Language Processing
TF-IDF	Term Frequency- Inverse Document Frequency
URL	Uniform Resource Locater
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative

Chapter 1: INTRODUCTION

1.1. Background

In the last few years, social media platforms have evolved a lot. Some of the most popular social media sites are Facebook, Twitter, Youtube and so on. People love to spend more time on exploring social media sites while using the internet. People share their day-to-day life on social media. They express themselves more freely on social media without being afraid of anything. The beauty of the social media is that it has empowered every section of the society without distinguishing between rich and poor, the so-called upper caste and lower caste, elite and non-elite. Any important incidence that takes place in our country or society or localities, people love to share their thoughts and feelings on such incidences through social media posts. Many times, nowadays, even the investigating agencies depend upon social media posts and analysis to solve criminal cases. The power of social media is such that political parties and leaders around the world create a special IT (Information Technology) team to keep an eye on what people are talking and thinking about the policies of a particular party or leader. Due to importance of social media, all news portals have made their strong presence on various social media platforms. News portals have huge potential to influence the thought process of individuals so the analysis of their headlines can give vital information regarding the events happening within the country and around the world.

As the world moves towards digitization, more digital data will be generated and those data would be vital to understand sentiment of people for positive cause, if used wisely. The misuse of those data could even lead to manipulation of feeling of the people towards some sensitive issues like national security.

1.1. Sentiment analysis

Sentiment analysis is considered as one of the important aspects in many areas like treatment of mental health problem, understanding the sentiment of people regarding the product of any companies, prediction of share markets, views of people regarding the launch of any scheme by the government, detecting fake news being circulated among the people, and so on. There are various ways in which one can try to understand the sentiment of people like conducting survey, preparing questionnaire, using the audio or video of the individuals or using the posts made by them on various social media sites.

Sentiment analysis is open research field. Various researches have been done and many are research works are going on regarding analysis. All researches offer something different than other but none of them are have been able to offer a complete solution and this is also due to the limitations of research in the field of Natural Language Processing and unavailability of open

and good datasets in the regard of this field. Sentiment analysis can be done through datasets available in various forms such as text, audio, video, images and so on. Comparatively it is easier to perform analysis on textual data due to the limitations of hardware and other resources for training a model. Sentiment analysis is an important field of NLP. NLP plays vital role in understanding the context of textual data. Also, it is very effective to unearth the pattern within unstructured data. After going through a number of research papers, the use of Bi-LSTM is found to be more effective as it takes context into consideration which plays significant role in improving the efficiency.

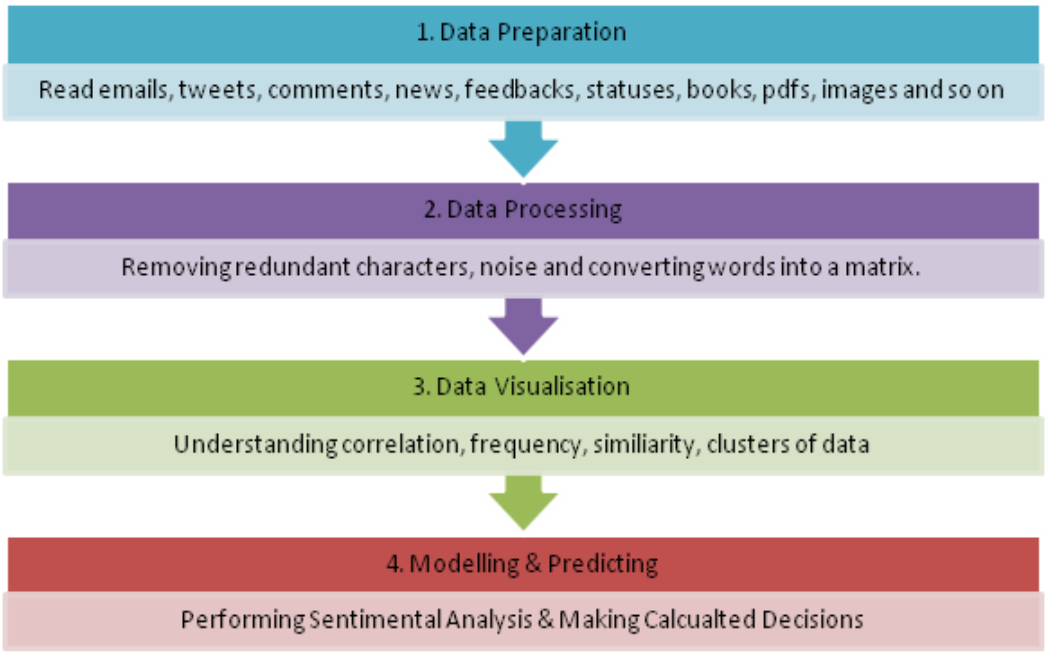


Fig 1.1 Four stages of sentiment analysis

Challenges of sentiment analysis:

- Lack of availability of enough domain specific dataset
- Multi-class classification of the data
- Difficult to extract the context of the post made by the users

1.2. Topic Modeling

Topic modeling is an unsupervised technique as it is used to perform analysis on text data having no label attached to it. As the name suggests, it is used to discover number of topics within the given sets of text like tweets, books, articles and so on. Each topic consists of consists of words where order of the words does not matter. It performs automatic clustering of words which best describes a set of documents. It gives us insight into number of issues or features users are taking about as a group.

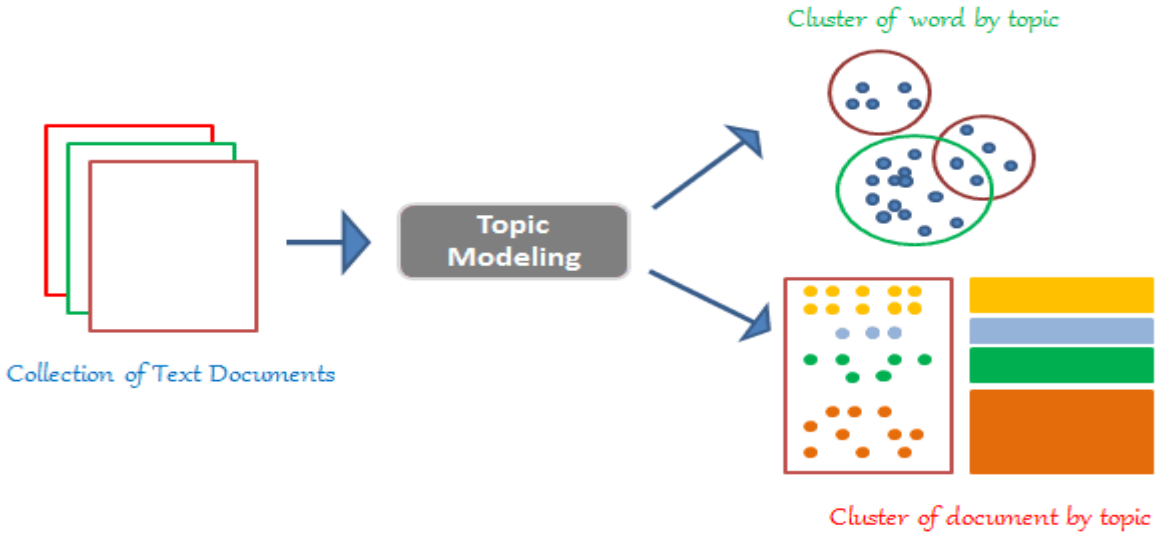


Fig 1.2 Topic Modeling

For example, let us say a company has launched a software in the market and it receives a number feedbacks regarding various features of the product within a specified time period. Now, rather than going through each review one by one, if we apply topic modeling, we will come to know how users have perceived the various features of the product very quickly. It is one of the important techniques to perform text analysis on unstructured data. After performing topic modeling, we can even perform topic classification to predict under which topic the upcoming reviews fall. There are various techniques to perform topic modeling, among which LDA is considered to be the most effective one.

1.3. Problem definition

Since there is no laboratory test available that could be used to predict the sentiment of people, news portal could offer a reliable platform for social science research. Due to the digitization of news portal, its outreach has increased and has the capacity to influence big events happening around the world due to which news portal has become a good source, for research in various fields like psychology, politics, entertainment and so on. Especially the news headlines are quite efficient for task like sentiment analysis due to limited number of number characters and targeted words used while publishing headlines so it would be useful to capture the emotions of people more efficiently and also easy to find topics being discussed in a collection of headlines. Both supervised and unsupervised techniques can be used to perform complete analysis of texts from various perspectives.

1.4. Objectives

- To quantitatively analyze the effect of various preprocessing techniques on the analysis of data.
- To perform analysis on textual data using both supervised and unsupervised techniques i.e., Bi-LSTM and LDA respectively for sentiment analysis and topic modeling.
- To perform exploratory analysis to get insights into useful information from the data.

Chapter 2: LITERATURE REVIEW

With the advancement and growth of online businesses, it has become easy for the companies to get feedback from the customers directly and work on their weaknesses to be able to strengthen their position in the market. They can also understand the emotions of people regarding their products through sentiment analysis [1]. Though the social media platforms have provided freedom of expression to people, it has led to cause some disturbing events as well. The hate speech over social media is one of the major concerns for various users. It has caused serious threat to the life of people. Sentiment analysis can play a big role in identifying hate speech over social media platforms and punch the perpetrators of those hate speeches. For this various deep learning techniques like CNN, RNN can be used [2].

Research conducted by Stanford University has used the emotion symbols to extract the tweets (using twitter's API) where each tweets represent either positive or negative sentiments. Various data preprocessing techniques have been applied to improve the accuracy of the classification models. Their research can be used to automatically classify the tweets without manual intervention. They have performed the overall sentiment analysis without focusing on a particular domain [3]. Twitter has become an important means to analyze the opinions, attitudes, emotional feeling of the people due to its growing popularity and influence among every class of people. There are many ways in which the sentiment analysis can be performed but the accuracy of the system is still a matter of concern due to lack of availability of proper datasets and the complexity involved withing Natural Language Processing (NLP) [4].

Sentiment analysis has many applications from entertainment industry to healthcare. A lot of researches are going on in this regard. Due to lack of availability of enough data, research work in this field has been hampered to some extent. Even if the data is available, it requires a lot of paper work and agreements to be fulfilled in order to obtain it. In fact, in some cases it could take few months as well in order to obtain the data [5].

With the advent of social networking sites, people love to share their feeling through it rather than discussing among the family members. Due to this amount of information available, many researches are being conducted in the field of mental health using the techniques of Natural Language Processing (NLP) [6]. The data related to mental health are still considered very confidential and not made publicly available by most of the authors due to agreement with the source from where the data have been collected. But with the invent of latest technologies we have now few APIs like tweepy, twitter 4j etc. which could be used to get data online to help analyze the sentiment of people using various social media sites. There are various social media websites which not only help to share written posts but also to share or host images like flickr, Instagram and so on. Even using and analyzing the pictures shared by different users on these platforms, to express their mood and feeling, can be used to understand their emotional side and sentiments [7]. Various classification techniques can be used to compare the result of classifying the data.

Though social media is often heard to contribute to mental illness in individuals due to increased anxiety, the same social media has also empowered people to overcome any stigma and outdated beliefs and express their thoughts without any fear. Due to this openness and freedom provided by social media, it becomes the duty of those platforms to do their bits in helping deal with existing and increasing mental health problems. Still, none of the researches have assured that it can help to accurately predict all the suicidal and non-suicidal posts made through social media by different users [8]. It is due to the problems associated with Natural Language Processing. Though it is difficult to provide 100% accurate result, it does give some sorts of idea and in some cases also helps to avert any kind of mishaps from happening. Except social media posts, there are also some other ways that could contribute to sentiment analysis of individuals. One of them could be classification of suicidal notes. It involves linguistic analysis. Different people express their feeling differently in the suicide notes. The use of language, pronouns, grammars vary vastly which makes it difficult to correctly classify suicidal and non-suicidal notes [9]. Some other researches have also been conducted to identify genuine and fake suicide notes. From traditional to latest machine learning techniques, a lot have been used in this problem domain. One of the reasons for reduced number of research in this domain is availability of small sample size, if in case any available. Use of NLP and text mining in sentiment analysis is still in its infancy, that's why we could see that the big social media tech giants find it difficult to accurately identify posts expressing negative sentiments [10].

In the early days of sentiment analysis, classifications were done based on the individual words rather than understanding the context in which those words have occurred [11]. Later on, wordnet based approach was proposed for sentiment analysis by calculating the distance between the word appeared in a text to the word "good" or "bad" [12]. Some researchers even used Cosine distance for better accuracy. The use of Bi-LSTM in sentiment analysis takes the context into consideration due to which it has been found to be more effective in sentiment analysis over other methods [13].

It is not easy to deal with huge text and get insights into what the text is trying to depict, topic modeling is one of the techniques to easily under the subjects depicted by the large collection of text [14]. Topic modeling has been studied and applied in various fields like political science, customer reviews, software engineering, medical, linguistic science [15]. Though there are many topic modeling algorithms available, it is necessary to perform the tuning and optimization of such algorithms to get reliable result. It is necessary to understand the underlying process in topic modeling algorithms to decide which algorithm best suit the given purpose [16].

Chapter 3: METHODOLOGY

In this work, we intend to apply supervised and unsupervised technique on a given dataset i.e., "times_of_india_headlines". The dataset is a collection of news headlines of one of the leading newspapers of India i.e., Times of India. For supervised learning Bi-LSTM (Bi-directional Long Short-Term Memory) will be used for performing sentiment analysis. Once the result from Bi-LSTM is obtained, it will be compared with ensemble model to see which performs better. LDA (Latent Dirichlet Allocation) will be used for topic modeling, which is an unsupervised technique. One of the major aspects of this work also includes quantitative analysis of data preprocessing, which help us to understand how data preprocessing plays vital role in obtaining reliable results, when passed through a model.

3.1. Overview of Methodology

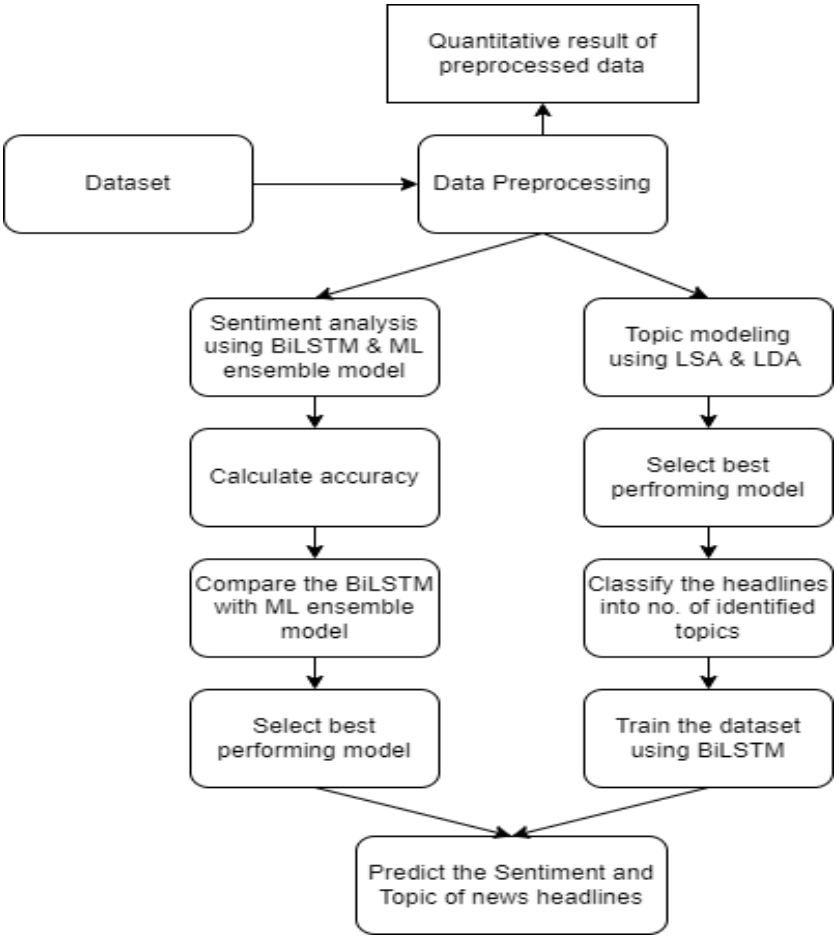


Fig 3.1 Proposed Methodology

In this process, the dataset of news headlines of “Times of India” will be collected. The collected data may not be in suitable form to be used directly for our model, so various steps like removing symbols, converting into small letters etc. would be applied for cleaning of data. Some quantitative analysis of preprocessed data would be performed to see the status of data before and after preprocessing. For sentiment analysis, ensemble model would be built using various machine learning models like ridge classifier, naïve bayes, ada boost and so on. Also, deep learning method i.e., Bi-LSTM would be used which is supposed to be more suitable for text analysis. The accuracy of both the model would be compared and the best model would be used for classification.

Similarly, for topic modeling two best known models i.e., LSA and LDA would be used to find the ideal number of topics hidden within the dataset. For this, the result would be observed for varying number of topics for both the models. The best performing model would be selected for classifying the dataset with the identified number of topics. Finally, Bi-LSTM would be used to train the dataset to classify the news headlines into various topics.

3.2. Dataset used

The dataset that will be used for this thesis is “times_of_india_headlines. This dataset contains a number of collection of headlines of the year 2019 AD. The dataset mainly covers the news of events in India. It has three columns and over 60 thousand of rows. As the dataset is not cleansed, it requires preprocessing task to be performed before applying it on model.

The columns with the datasets are listed below:

- Text (i.e., news headlines)
- Published_date
- Sentiment

text	published_date	Sentiment
MP to study new Motor Act before implementing ...	2019-09-03	Negative
Bhujbal says higher power rate driving industr...	2019-12-10	Negative
Mandira Bedi shares a monochrome picture in a ...	2019-12-12	Negative
Harassment in office: DM seeks action-taken re...	2019-11-18	Negative
Two students end their lives	2019-06-09	Negative
...	...	...
Going to India Gate tonight? You could be in f...	2019-12-31	Positive
40;000 cops to ensure safe New Year celebratio...	2019-12-31	Positive
Ten more private trains from Mumbai on cards	2019-12-31	Positive
A trip to a tribal village with super cop Vija...	2019-12-31	Positive
Wi-Fi at over 5;500 railway stations; thanks t...	2019-12-31	Positive

Fig 3.2 Sample of dataset

3.3 Data preprocessing

The dataset in original form cannot be fed directly as input to the model as it contains lots of redundant data. Mainly the data collected from online platforms are either unstructured or semi-structured as they may contain such information which may be insignificant for our analysis. So, we need to remove those insignificant data. In our case the headlines may contain URLs, special symbols, white spaces, words of different case and so on. Proper cleaning of data can help us reduce the number of features extracted from the dataset. Also, because our dataset is large, it can reduce the training time and helps in performing precise prediction.

Some major steps involve in data preprocessing are listed below:

• converting the tweets to lower case letters

• removing stop words

• removing the URLS that could be present

• removing emoticons

• removing non-ascii characters

• removing non-English tweets

• removing special characters

• stemming down the words

	sentiment	text
0	0	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	is upset that he can't update his Facebook by ...
2	0	@Kenichan I dived many times for the ball. Man...
3	0	my whole body feels itchy and like its on fire
4	0	@nationwideclass no, it's not behaving at all....

	text	target
0	- aww, that's a bummer. you shoulda got dav...	0
1	is upset that he can't update his facebook by ...	0
2	i dived many times for the ball. managed to s...	0
3	my whole body feels itchy and like its on fire	0
4	no, it's not behaving at all. i'm mad. why am...	0

Fig 3.3 Sample of preprocessed data

3.4. Word Embedding

It is difficult to perform analysis on the text so we use word embedding techniques to convert the texts into numerical representation. This is also called vectorization of the words. It is a representation technique for text in which words having same meaning are given similar representation. It helps to extract features from the text.

i. Bag of Words (BoW)

It is one of the most widely used technique for word embedding. This method focuses on the frequency of the words. It is easy to implement and also efficient at the same time. In Python, we can use CountVectorizer() function from sklearn library to implement Bag of Words. A graphical representation of bag of Words has been shown in the figure below.

The reason for selecting BoW as word embedding technique is that it is still widely used technique due to its simplicity in implementation. It is also very useful when working on a domain specific dataset. If not entirely, it does give some idea to the researchers regarding the performance of the work.

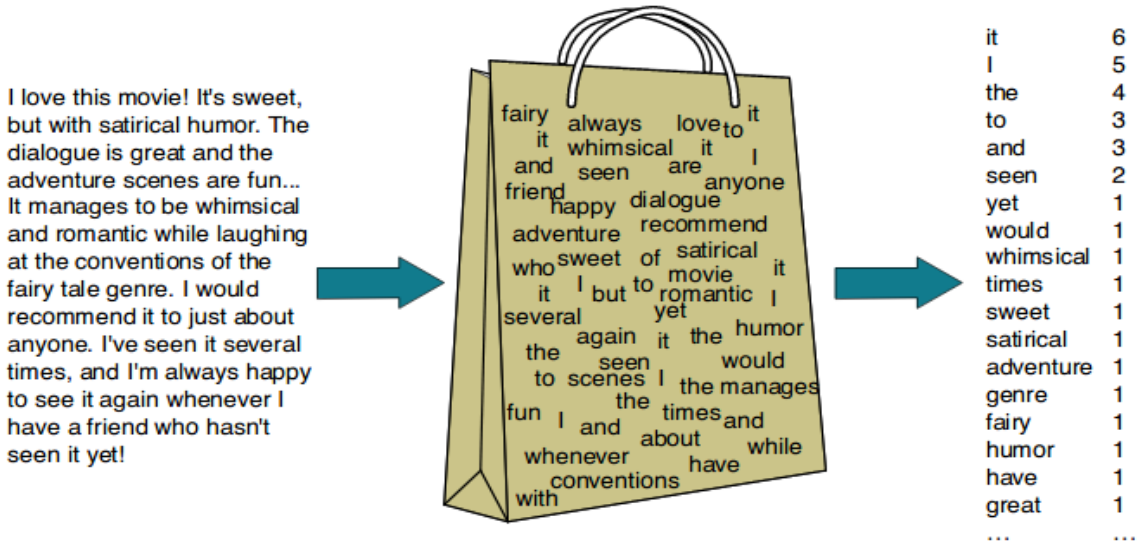


Fig 3.4 Representation of BoW

ii. TF-IDF (Term Frequency- Inverse Document Frequency)

Term Frequency is used to count the number of words in a document. It is not necessary that the words with higher frequencies tend to represent significant information about the document. A word appearing in a document for many times does not mean it is relevant and significant all the time. Many times, words with less frequencies carry more significant meanings about the document. One way to normalize the frequency of words is to use TF-IDF. Inverse Document Frequency (IDF) is used to calculate the significance of rare words or words with less frequencies.

$$TF(i,j)=n(i,j)/\Sigma\, n(i,j)$$

Where,

$n(i,j)$= number of times nth word occurred in a document
$\Sigma n(i,j)$ = total number of words in a document.

$$IDF=1+\log(N/dN)$$

Where

N=Total number of documents in the dataset
dN=total number of documents in which nth word occur

The TF-IDF is obtained by

$$TF\text{-}IDF=TF*IDF$$

The reason for using TF-IDF is it can be used to find and remove stop words in the textual data. It helps to find unique identifier in a text. It helps to understand the importance of a words in entire document which in turn can help in text summarization.

3.5. N-grams

N-gram is N sequence of words. It can be unigram (one word), bigram (sequence of two words), trigram (sequence of three words) and so on. It focuses on sequence of words. Such method is very useful in speech recognition, predicting input text and so on. It helps us to predict the next words that could occur in a given sequence. Search engines also uses n-gram technique to predict the next word while search query is typed in a search bar.

Let us consider a sentence:

“This is so weird.”

The bigram of this sentence can be written as:

- “This is”
- “is so”
- “so weird”

With n-grams, we can use a bag of n-grams (for example, bag of bigrams) instead of only using a bag of words. A bag of bigrams or trigrams is more powerful that using just bag of words as it takes the context into consideration as well.

3.6 Sentiment analysis

3.6.1 RNN

In traditional neural networks, inputs and outputs were independent of each other. To predict next word in a sentence, it is difficult for such model to give correct output, as previous words are required to be remembered to predict the next word. For example, to predict the ending of a movie, it depends on how much one has already watched the movie and what contexts have arrived to that point of time in the movie. In the same way, RNN remembers everything. It overcomes the shortcomings of traditional neural network with the help of hidden layers. Because of the quality to remember the previous inputs, it useful in prediction of time series. This is called Long Short-Term Memory (LSTM).

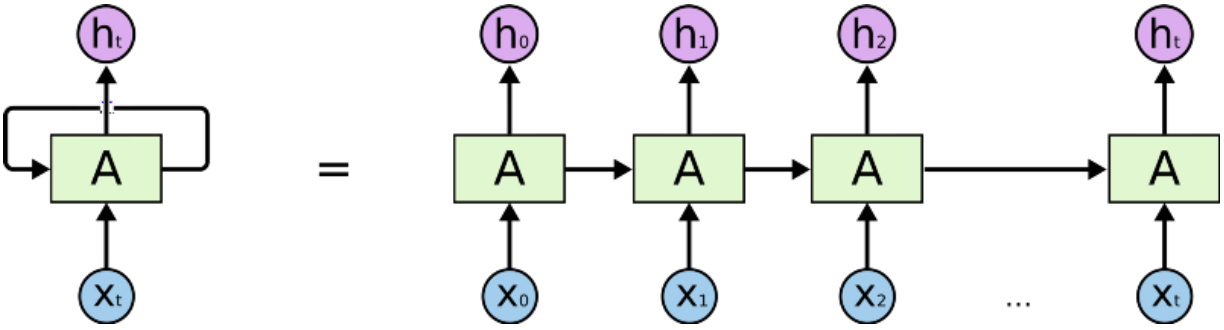


Fig 3.5 LSTM Architecture

Calculation of current state:

$$h_t = f(h_{t-1}, x_t)$$

Where,

h_t = current state
h_{t-1} = previous state
x_t = input state

formula for activation function (tanh):

$$h_t = \tanh (W_{hh}h_{t-1} + W_{xh}x_t)$$

w_{hh} = weight at recurrent neuron
w_{xh} = weight at input neuron

formula for output:

$$y_t = W_{hy}h_t$$

Where,

Y_t = output

W_{hy} = weight at output layer

3.6.2 Bi-LSTM

Combining two independent RNN together forms a Bi-LSTM (Bidirectional Long Short-Term Memory). It allows the network to have both forward and backward information. Bi-LSTM gives better result as it takes the context into consideration.

Example:

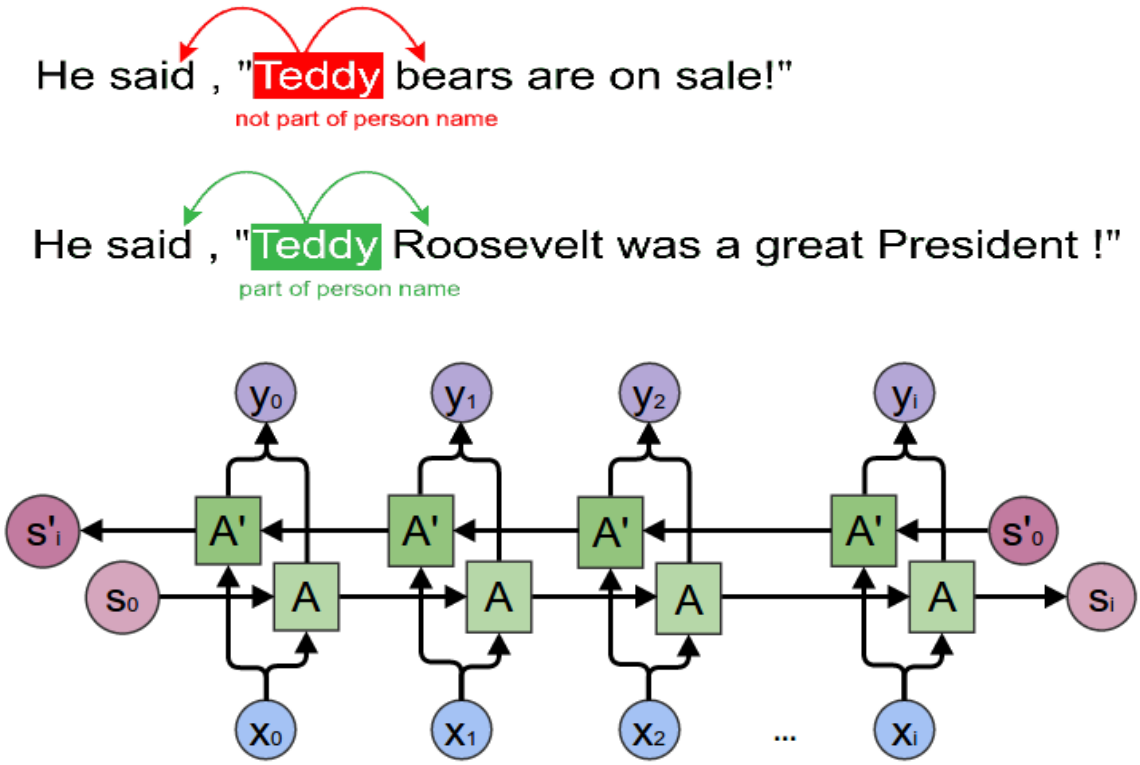


Fig 3.6 Bi-LSTM Architecture

In NLP, to find the next word, it is not that we need only the previous word but also the upcoming word. As shown in the example above, to predict what comes after the word "Teddy", we can't be dependent only upon the previous word (which is "said" in this case) because it same in both the cases. Here, we need to consider the context as well. If we predict "Roosevelt" in the first case then it will be out of context as it will read as [He said, "Teddy Roosevelt are on sale!"]. So, Bi-LSTM is important to taken context into consideration.

3.6.3 Ensemble model

Various machine learning techniques will be to train the dataset for sentiment analysis and their accuracy will be calculated. Based on their calculated accuracy, top performing models will be selected to build an ensemble model to get a single accuracy. This obtained accuracy will be compared with the accuracy obtained using Bi-LSTM to find which performs better. Voting classifier will be used to build an ensemble model.

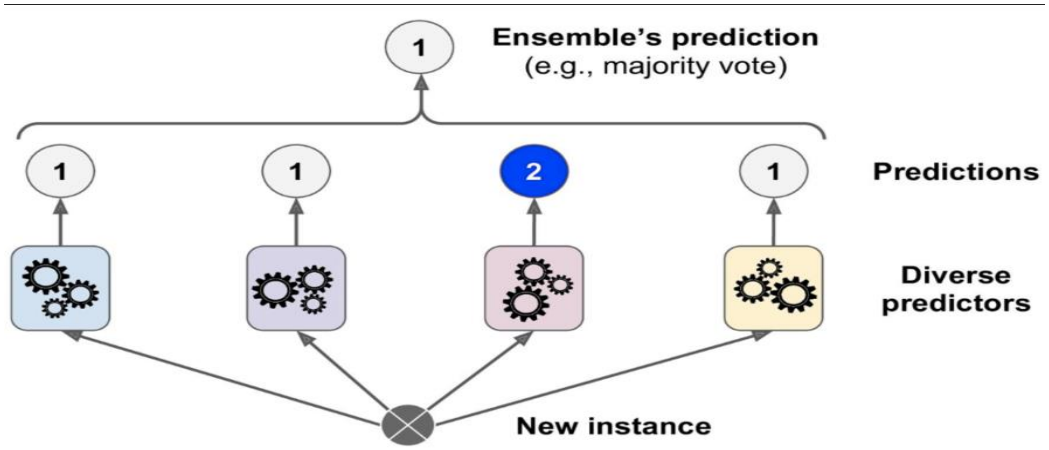


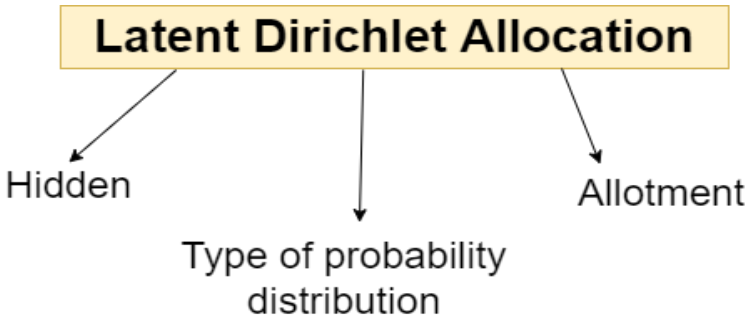
Fig 3.7 Voting classifier ensemble model

3.7 Topic Modeling

Topic modeling is an unsupervised method used to perform text analysis. When we are given large sets of unlabeled documents, it is very difficult to get an insight into the discussions upon which the documents are based upon. Here comes the role of topic modeling. It helps to identify a number of hidden topics within a set of documents. Based on those identified topics, the entire sets of documents can be classified. Also, it can be used to predict the topic of upcoming documents. It can have various applications like customer reviews, story genre prediction, tweets classification, news headlines classification and so on.

3.7.1 LDA

LDA (Latent Dirichlet Allocation) is one of the most used and well-known technique to perform topic modeling.



The word latent means hidden because we don't know what topics a set of documents contain. Based on probability distribution of occurrence of words in each document, they are allocated to defined topics.

Working of LDA:

- What we want for LDA is to learn topic mix in each document and also learn the word mix in each document.
- We choose a random number of topics for the given dataset.
- Assign each word in each given document to one of the defined topics, randomly.
- Now, we go through each and every word and to which topic those words are assigned to in each of document. Then, it is analyzed how often the topic occurs in the document and how often the word occurs in the topic as a whole. Base on this analysis, new topic is assigned to the given word.
- It goes through a number of such iterations and finally the topic will start making sense. We can analyze those found topics and assign a suitable name to those topics which best describe them.

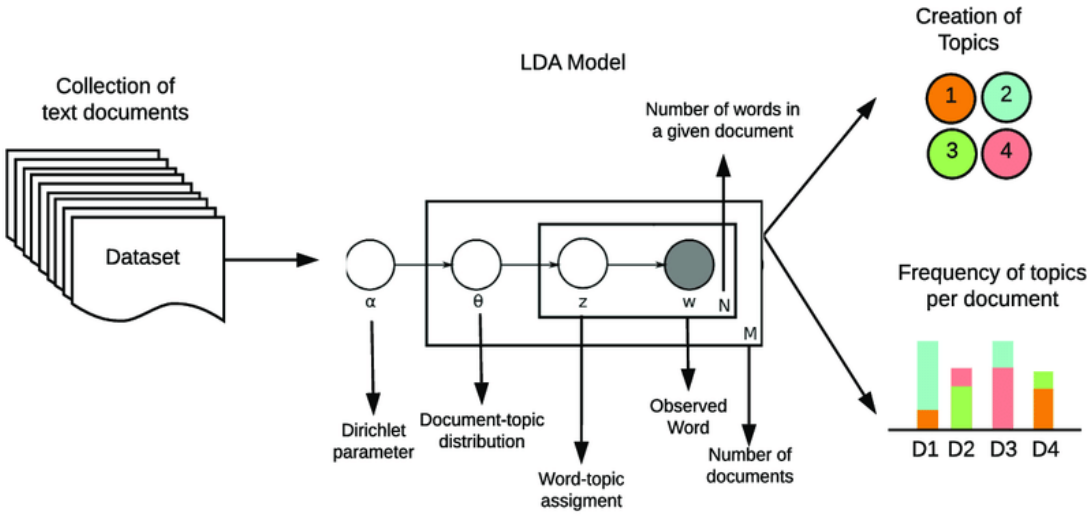


Fig 3.8 Schematic diagram of LDA algorithm

3.7.2 LSA

LSA (Latent Semantic Analysis) is another technique used for topic modeling. The main concept behind topic modeling is that the meaning behind any document is based on some latent variables so we use various topic modeling techniques to unravel those hidden variables i.e., topics, so that we can make sense out of given document. LSA is mostly suitable for large sets of documents. It converts the documents into document term matrix before actually deriving topics from the documents.

Working of LSA:

- The given text is converted into document- term matrix using either bag of words or Term Frequency- Inverse Document Frequency.
- Then, using Truncated Singular Value Decomposition (SVD). It is at this stage the topics within the documents are identified. Mathematically, it can be given as,

$$A \approx U_t S_t V_t^T$$

Though it may look difficult to understand at first glance, in simple terms what the above formula represents is that it simply decomposes high dimensional matrix into smaller matrices i.e., u, s and v, where,

A = n*m document-term matrix (n = no. of documents and m = no. of words)

U = n*r document-topic matrix (n = no. of documents and r = no. of topics)

S = r*r matrix (r = no. of topics)

V = m*r word-topic matrix (m = no. of words and r = no. of topics)

- Finally, we can now classify which document belongs to which topics.

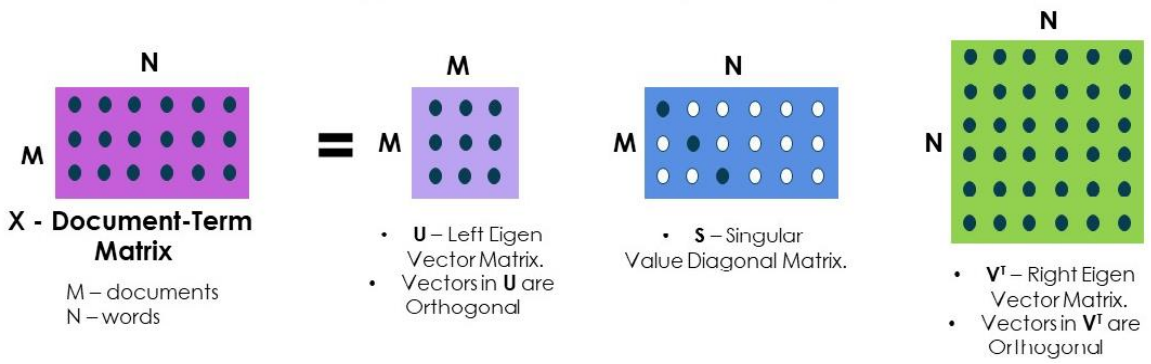


Fig 3.9 Schematic diagram of LSA algorithm

3.8 Confusion Matrix

Confusion matrix is an important metric to evaluate the performance of classifier. The performance of a classifier depends on their capability predict the class correctly against new or unseen data. It is one of the easiest metrics for finding the correctness and accuracy of the model. The confusion matrix in itself is not a performance measure but all the performance metrics are based on confusion matrix. The ideal situation for any classification model would be when FP=0 and FN=0 but that's not the case in real life. Depending upon the situation, we might want to minimize either FP or FN.

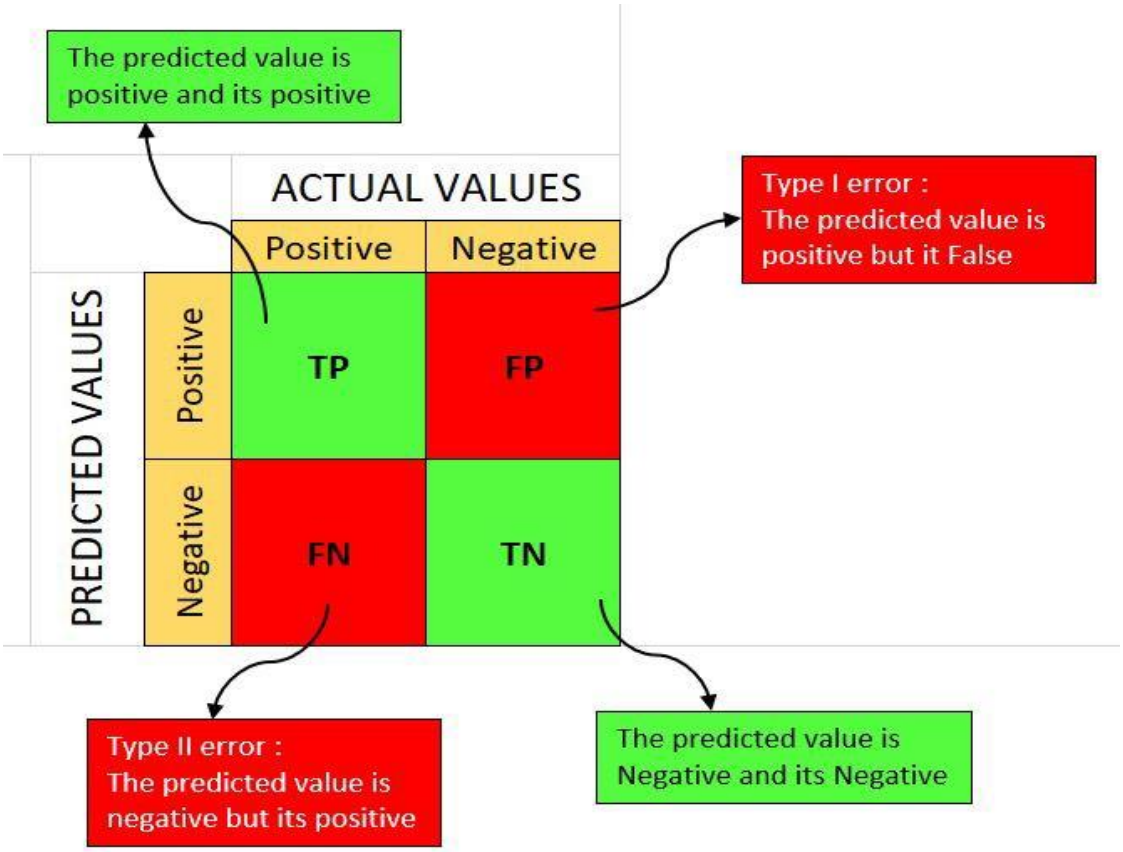


Fig 3.10 Confusion Matrix

Let us consider both the situations with the help of example.

Case 1: Minimizing FN

Let,

1 = person having cancer

0 = person not having cancer

In this case, having some False Positive (FP) cases might be okay because classifying a non-cancerous person as cancerous does not affect a lot because on further test we will anyway find out that the particular person does not have cancer. But having False Negative (FN) cases can be hazardous because classifying a cancerous person as non-cancerous can cause serious threat to the life of that person. So, in this case we need to minimize FN.

Case 2: Minimizing FP

Let,

1 = Email is a spam

0 = Email is not spam

In this situation, having False Positive (FP) cases i.e., wrongly classifying non-spam or important email as spam can cause serious damage to the business, financial loss to the individuals and so on. Thus, in this situation we need to minimize FP.

Various metrics based on confusion matrix

1. Accuracy: It is the number of correct predictions made by the model over all kinds of predictions made. It is a good measure when the target variable classes in the data are nearly balanced.

   Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

2. Precision: It is defined as the ratio of number of positive samples correctly classified as positive to the total number of samples classified as positive (either correctly or incorrectly). It reflects how reliable the model is in classifying the samples as positive. It is used if the problem is sensitive to classifying a sample as positive in general.

   Precision = $\frac{TP}{TP+FP}$

3. Recall or sensitivity: It is defined as the ratio of number of positive samples correctly classified as positive to the total number of actual positive samples. It measures the model's ability to detect positive samples. Higher the recall value, more positive samples are detected. It is independent of how the negative samples are classified. It is used if the goal is to detect all the positive samples without caring whether the negative samples would be misclassified as positive.

   Recall = $\frac{TP}{TP+FN}$

Chapter 4: RESULTS AND ANALYSIS

4.1 Preprocessed data

8	Negative	train cancel anna tuesday
9	Negative	polic pick trash murder inform dad
10	Negative	shortag psychiatrist tone punjab drug read drive
11	Negative	top point view adher human right polici
12	Negative	vijayawada doctor clash accid victim kin
13	Negative	al_caps_cpm want templ built write centr
14	Negative	delhi takh line bid regain martha strongman imag
15	Negative	ahmedabad onli eleph take part intra
16	Negative	cyclon bulbul pilot agar island help evacue
17	Negative	properti tax default owe core chandigarh civic bodi
31543	Positive	miner magic effect coat condit
31544	Positive	meekli look news may june
31545	Positive	deep cool
31546	Positive	to right everi
31547	Positive	start may
31548	Positive	watch new aerial nayaki soon televis
31549	Positive	skil music banarasi
31550	Positive	biryani treasur recip india
31551	Positive	vanya ramkumar debut opposit murat iowa
31552	Positive	lid fun day moss
31553	Positive	banana bring love angl canist movi
31554	Positive	judah parti love dog reveal

Fig 4.1 Sample of Preprocessed data

Some quantitative results of preprocessing are given below:

Total sentences: 63080

Total Words before preprocess: 738298
Total Distinct Tokens before preprocess: 151401
Average word/sentence before preprocess: 11.78589436

Total Words after preprocess: 435436
Total Distinct Tokens after preprocess: 46286
Average word/sentence after preprocess: 6.94530312
Total emoticons: 18413
Total slangs: 509

4.2 Results of Sentiment analysis

4.2.1 Sentiments distribution

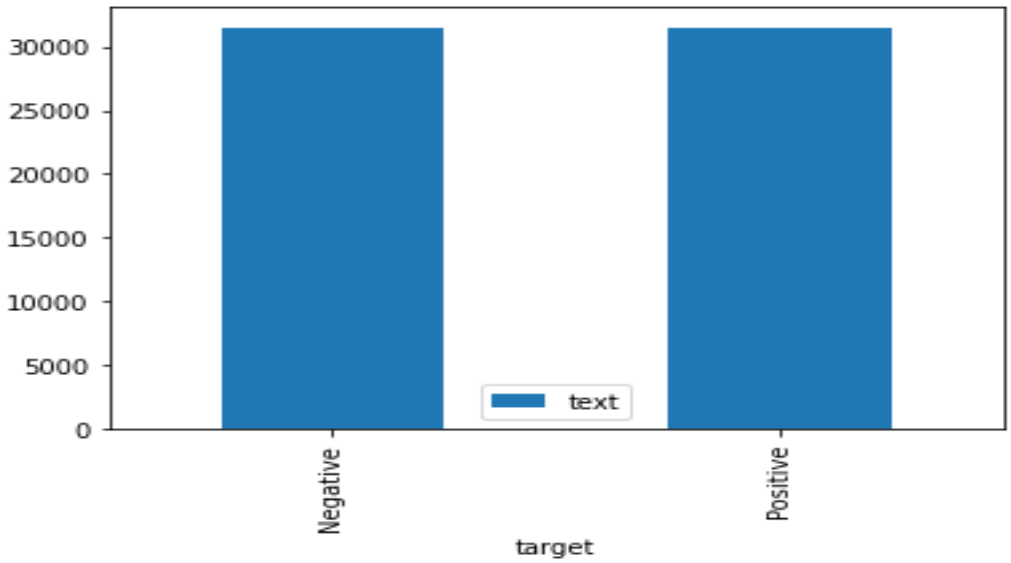


Fig 4.2 Sentiments distribution

As it can be seen in the fig 4.2, the positive and negative classes have been equally distributed. This helps to make our classification more accurate and improves the prediction of new data.

4.2.2 Distribution of Length of new headlines

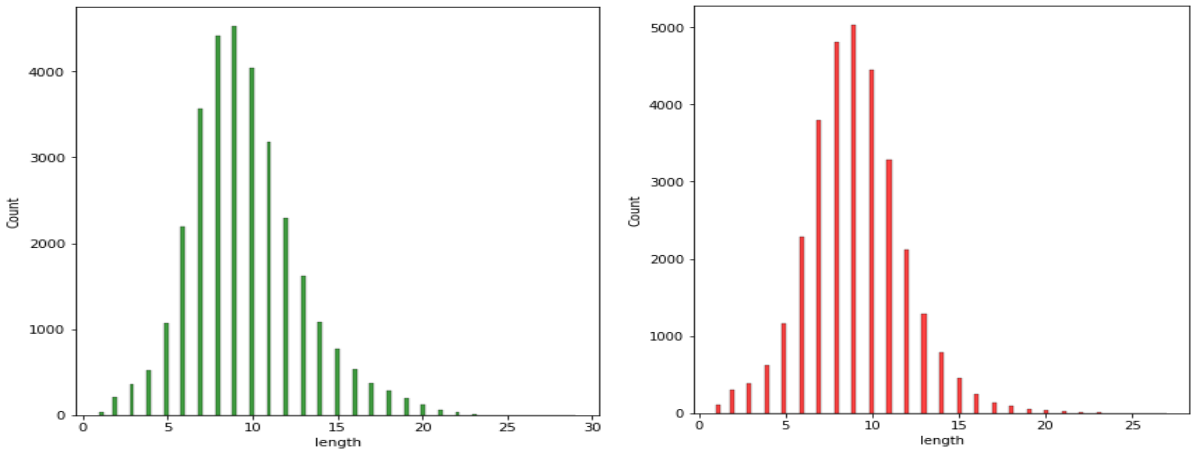


Fig 4.3 News length distribution

From the fig 4.3, it can be seen that in both the positive and negative cases of length distribution of news headlines, there is only one peak which means there is no mixing of other classes within these two classes. Had there been more than one peak in any of these classes, there could have been neutral headlines mixed within these classes but there is no such case as observed from the figure above.

4.2.3 Positive vs Negative frequency graph (for same word)

	negative	positive	t
rs	624	660	
man	427	478	
delhi	391	463	
held	418	452	
year	337	352	
lakh	267	321	
government	296	321	
woman	275	315	
road	316	314	
old	284	292	

Fig 4.4 Words occurrences in both positive and negative classes

The fig 4.5 shows the plot of words which have occurred in both the positive and negative sentiments. The negative frequency of such words plotted on X-axis and positive frequency plotted on Y-axis. It can be seen there are many words whose positive and negative frequency occurrences are below 200. There are very few words which have occurred in both the positive and negative sentiments with higher frequencies. Some words with occurrences in both the sentiments have been shown in fig 4.4 below.

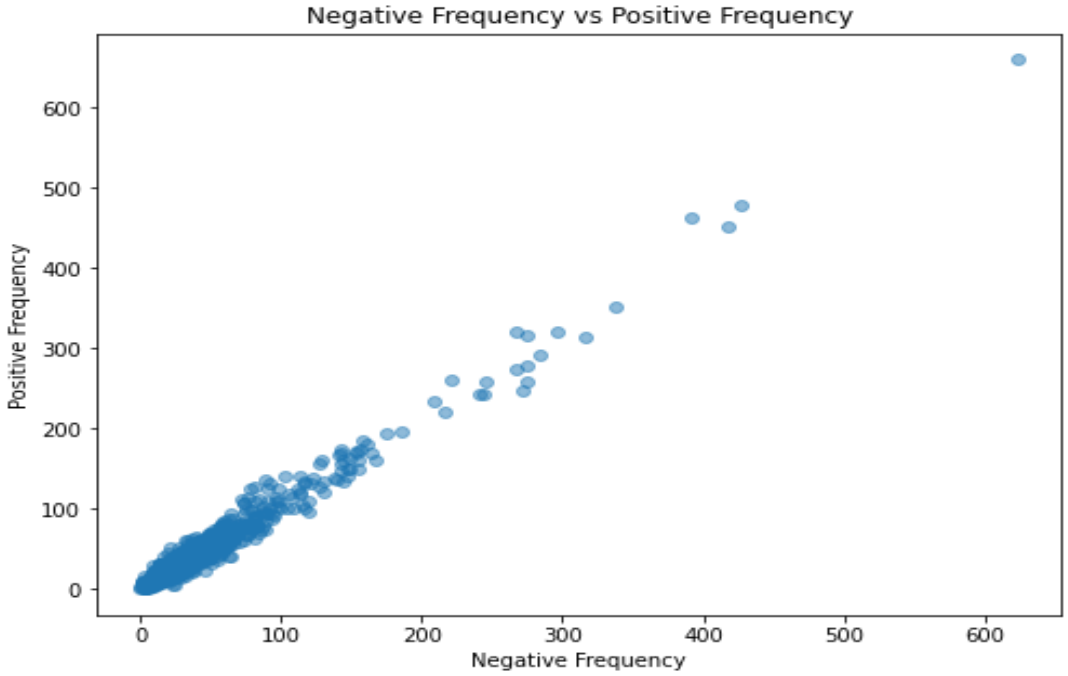


Fig 4.5 Positive vs Negative frequency graph

4.2.4 Accuracy with various n-grams (for ML model)

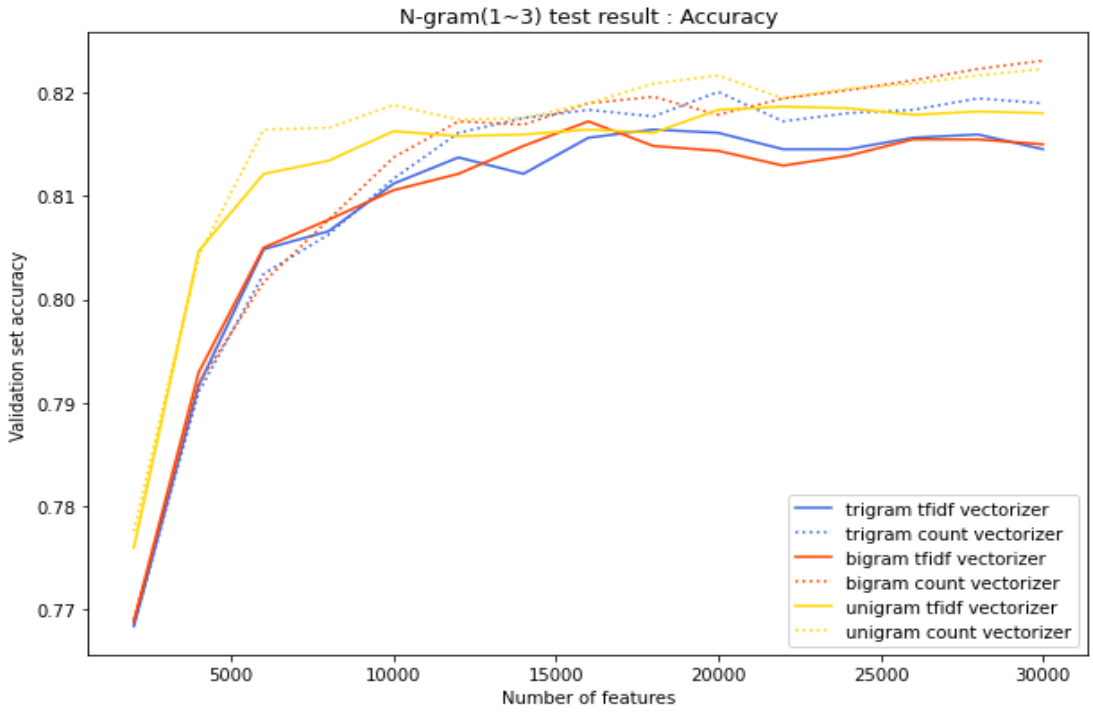


Fig 4.6 Accuracy graph for n-grams

The fig 4.6 shows the plot of n-grams with TF-IDF and count vectorizer word embedding techniques. The above figure can be used to compare the accuracy of word embeddings with various n-grams so that the one with highest accuracy can be used for training the classification model. In our case, it can be visualized that the accuracy is higher for unigram for both embedding techniques when the number of features is less than ten thousand, but as the number of features are increased the bigram count vectorizer is seen to have higher accuracy at thirty thousand features. The trigram TF-IDF and bigram TF-IDF is seen to have lower accuracy at thirty thousand features. Also, the difference between higher and lower accuracy at thirty thousand features is approximately one percent. Because there is not much difference between the accuracies of various n-grams, the trigram with TF-IDF word embedding would be taken for training the model. This is because trigram helps to take the context of the word into consideration. This helps in overcoming the difficulty with the words which have appeared in both positive and negative news headlines. Also, TF-IDF helps to give more importance to the rare words that have occurred in news headlines.

4.2.5 Accuracy of Ensemble model (ML algorithm)

After performing classification using various Machine Learning algorithms, accuracy of those algorithms was used to form an ensemble model using Voting classifier model. The ML models used were Logistic regression, Ridge classifier, Linear SVC, Ada boost, Passive aggressive classifier. The ensemble model can be of two types i.e., hard voting and soft voting. In hard voting, the sum of class predicted by each model is calculated and the class with highest number of votes is chosen as the final predicted class of the ensemble model. Similarly, in case of soft voting, the sum of predicted probabilities for each class made by the each model is calculated and the class with highest probability is chosen as the final predicted class.

The accuracy of ensemble model using both voting hard and voting soft are given below:

Accuracy with Voting hard = 81.67%

Accuracy with Voting soft = 80.45%

4.2.6 Model accuracy and loss graph for Sentiment analysis (using Bi-LSTM)

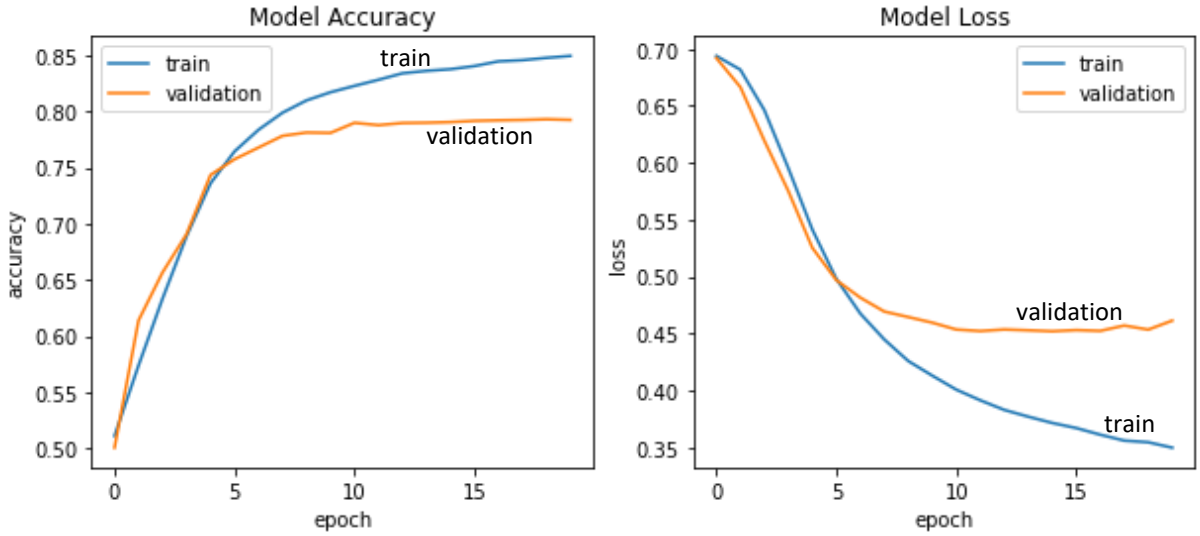


Fig 4.7 Accuracy and loss graph for sentiment analysis

The fig 4.7 shows Bi-LSTM model accuracy and loss graph for sentiment analysis. The callback function has been employed to automatically stop the training when the accuracy or loss condition is met. The maximum epochs that have been required in this case is 20. The accuracy for training set is higher than validation set. Also, the loss for training set is lower than that of validation set because the validation set is unseen data whereas as the training set is seen data, which in turn increases the accuracy of training set.

4.2.7 Results of Bi-LSTM model for sentiment analysis

No. of epochs = 20
Batch size = 64
Learning rate= 0.1

Precision = 75.38%
Recall or Sensitivity = 84.57%
Model Accuracy = 84.92%

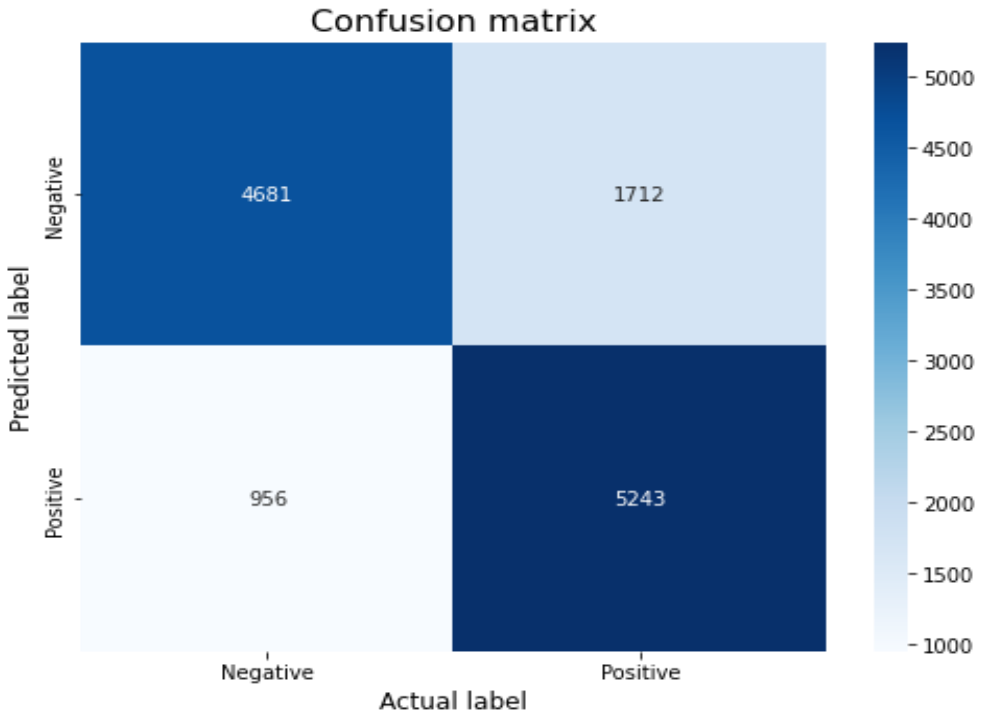


Fig 4.8 Confusion matrix for sentiment analysis (Bi-LSTM)

4.3 Results of Topic Modeling

4.3.1 LSA clustering graph for varying topics

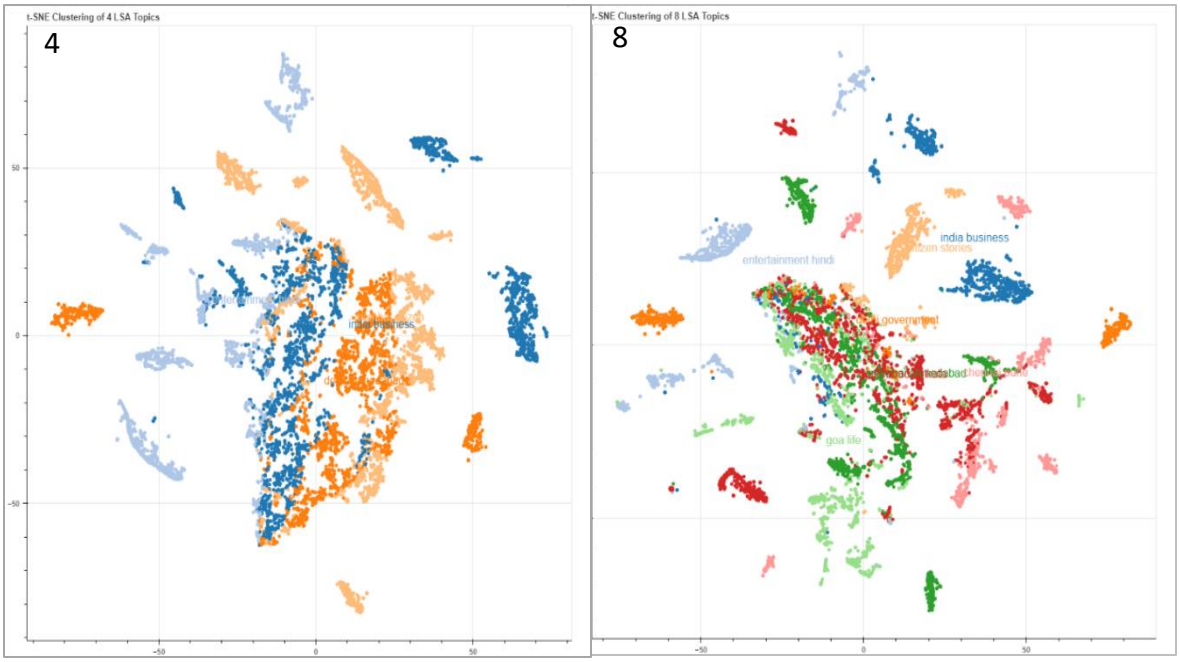


Fig 4.9 LSA topic modeling for 4 and 8 topics

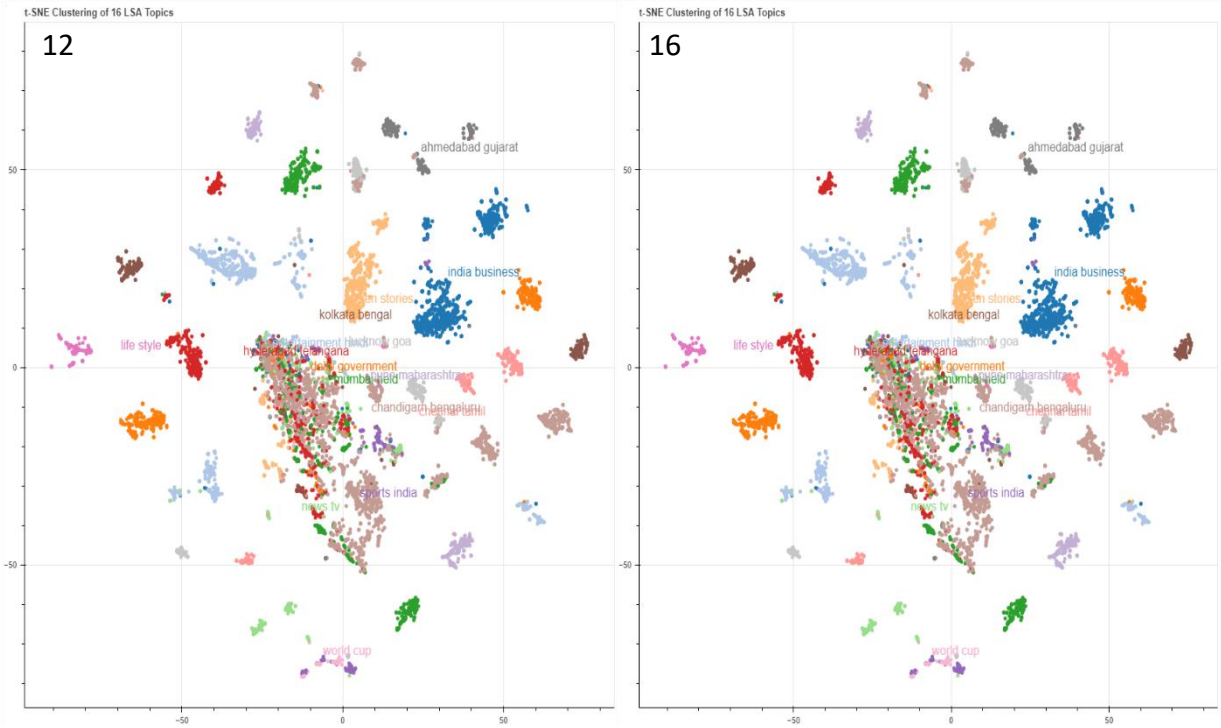


Fig 4.10 LSA topic modeling for 12 and 16 topics

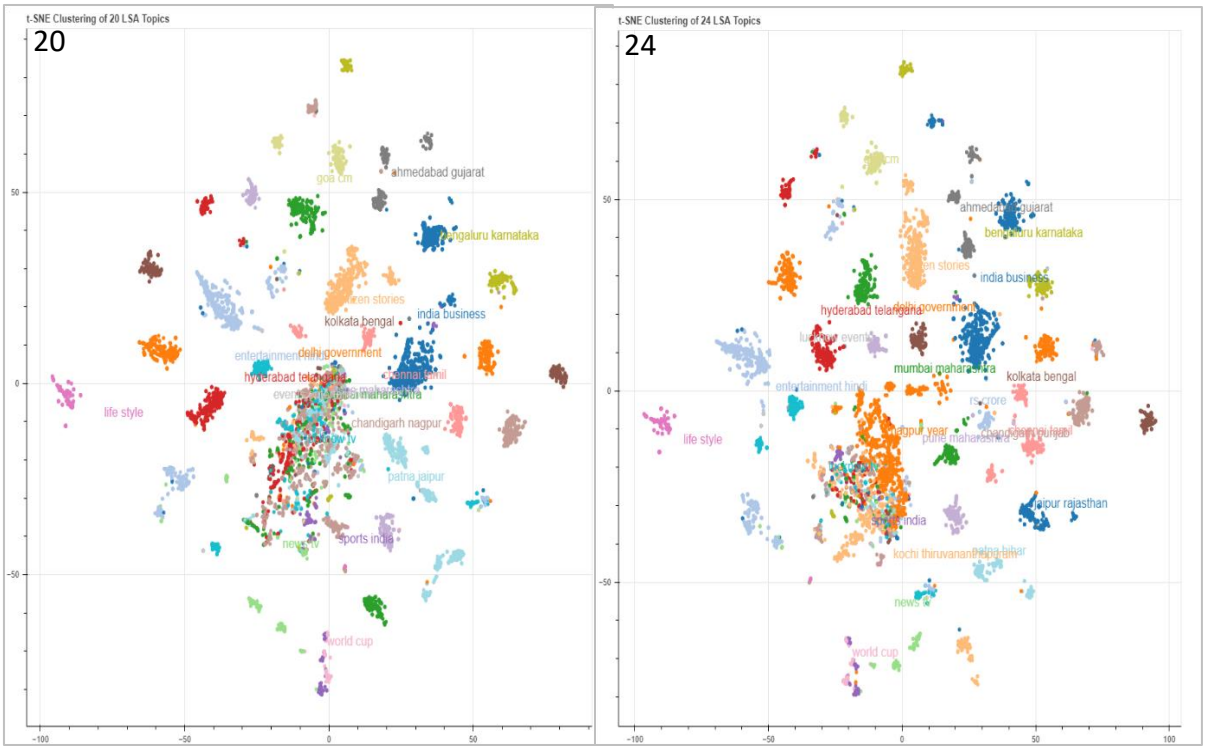


Fig 4.11 LSA topic modeling for 20 and 24 topics

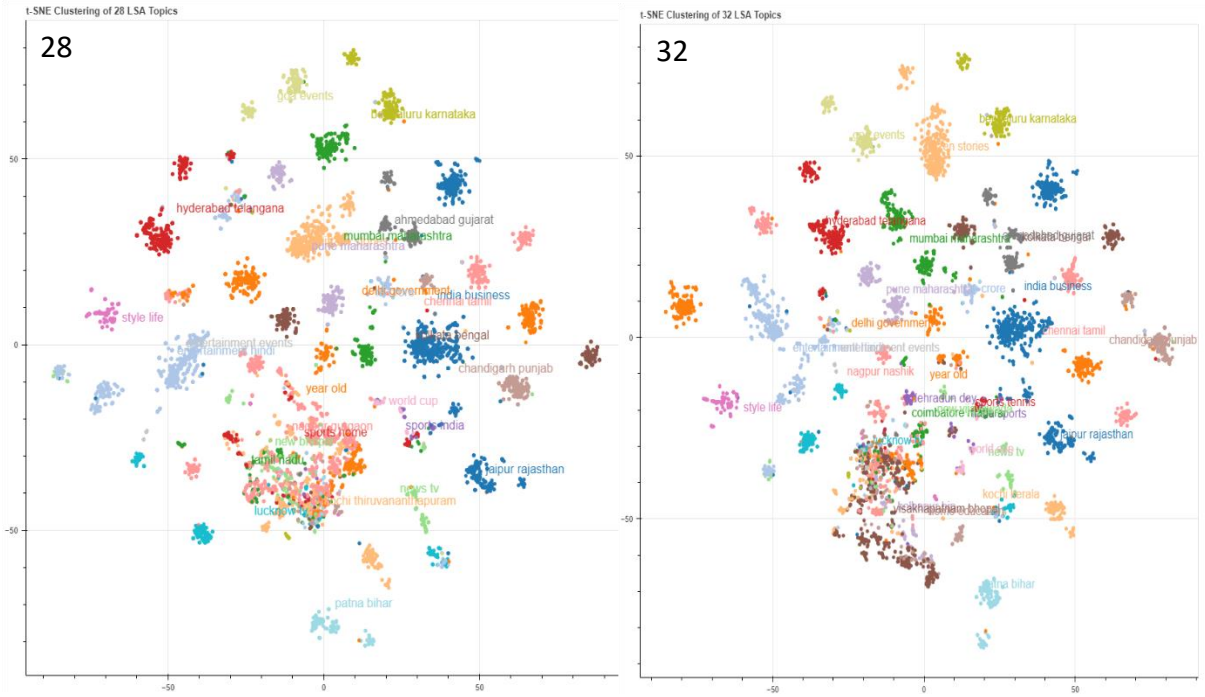


Fig 4.12 LSA topic modeling for 28 and 32 topics

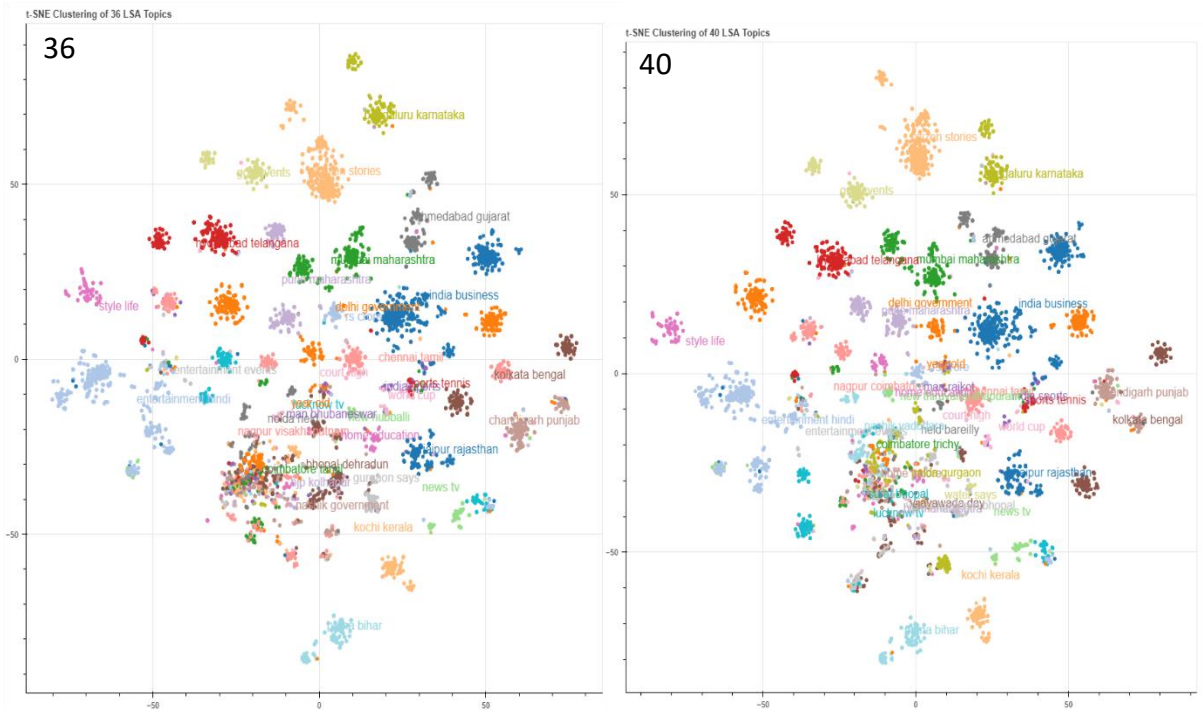


Fig 4.13 LSA topic modeling for 36 and 40 topics

4.3.2 LDA clustering graph for varying topics

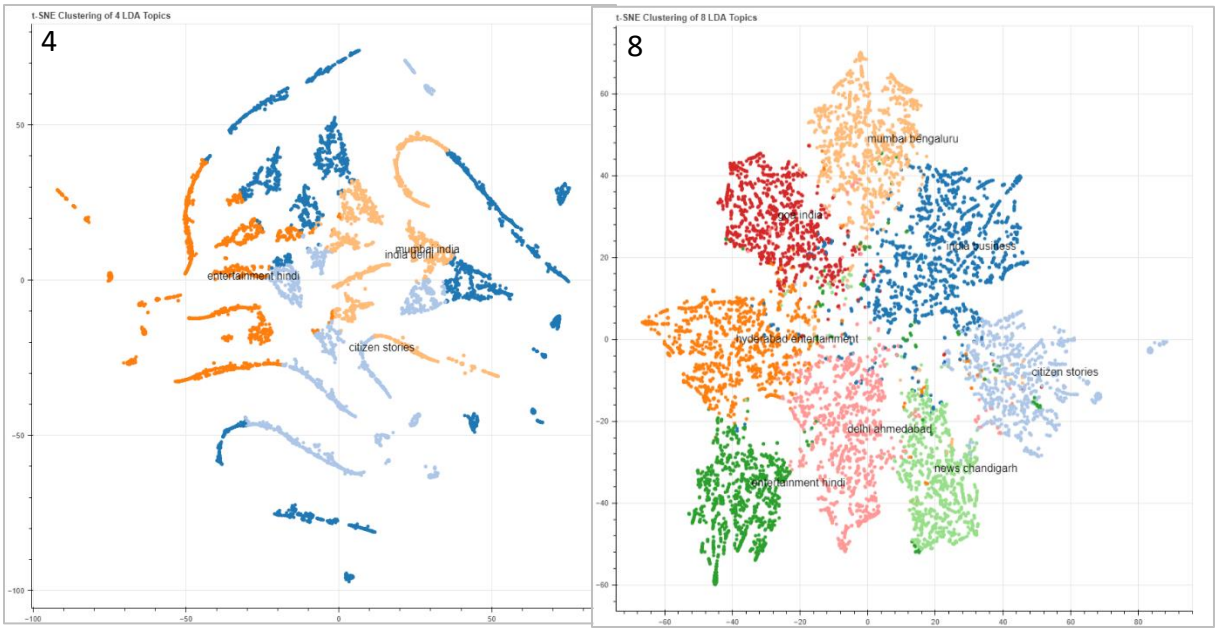


Fig 4.14 LDA topic modeling for 4 and 8 topics

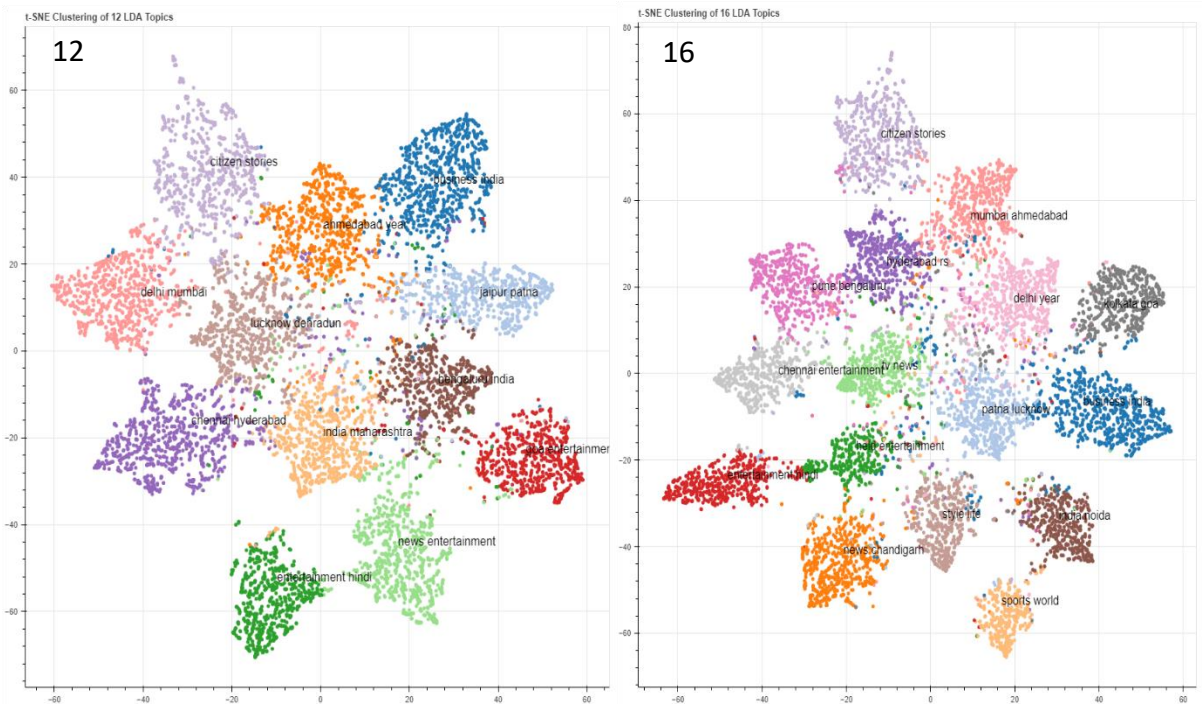


Fig 4.15 LDA topic modeling for 12 and 16 topics

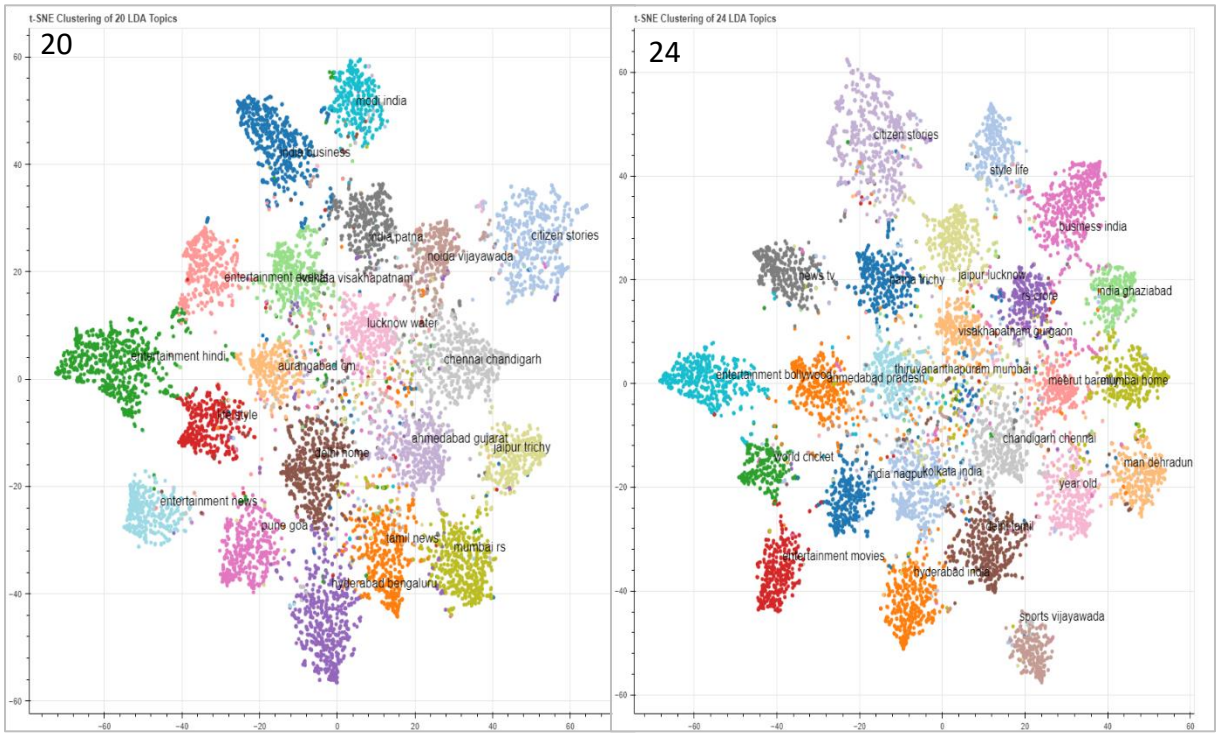


Fig 4.16 LDA topic modeling for 20 and 24 topics

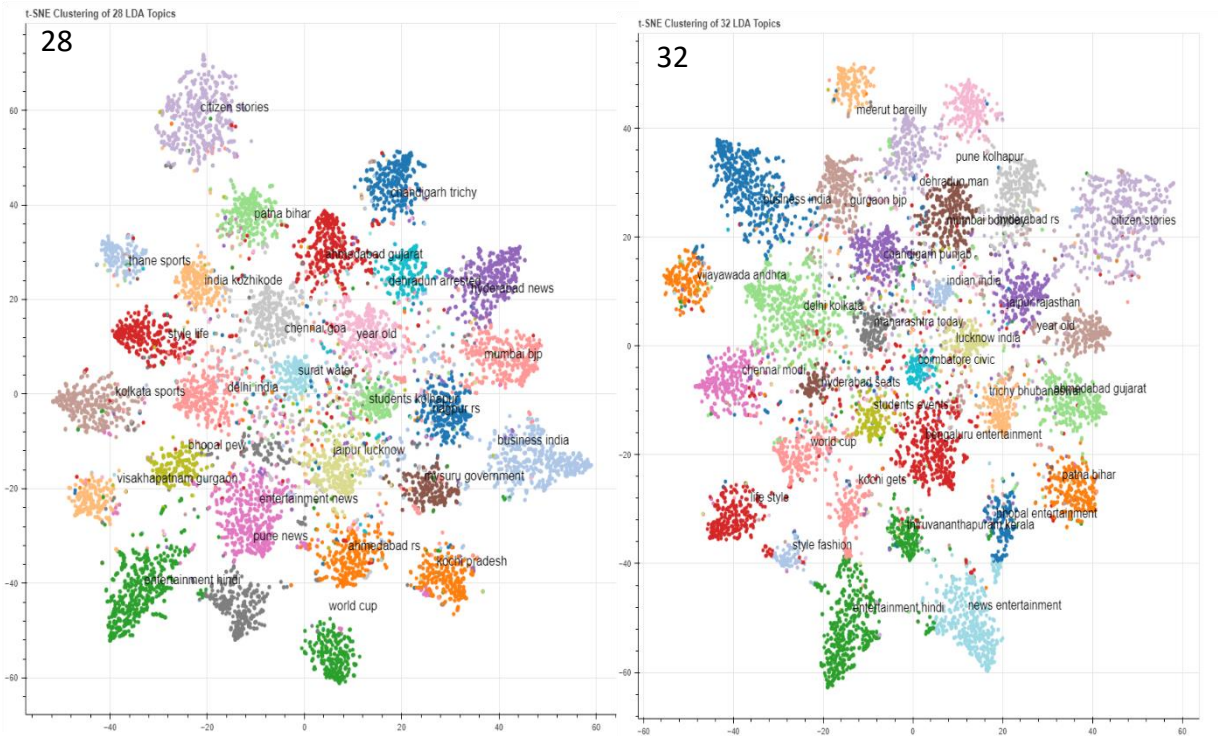


Fig 4.17 LDA topic modeling for 28 and 32 topics

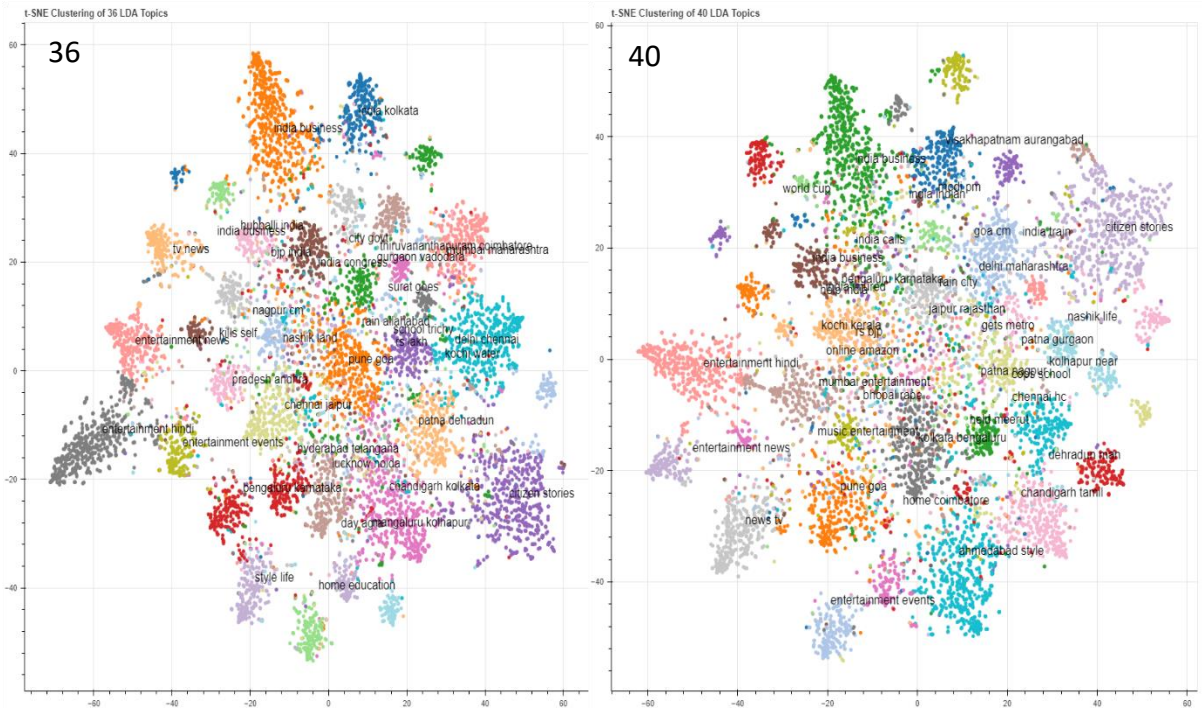


Fig 4.18 LDA topic modeling for 36 and 40 topics

4.3.3 LSA bar graph for topics distribution

Fig 4.19 LSA topic distribution for varying number of topics

4.3.4 LDA bar graph for topics distribution

Fig 4.20 LDA topic distribution for varying number of topics

Comparing the topics distribution graph for varying number of topics for both LSA and LDA, it can be observed that LDA has more uniform topic distribution than LSA. So, from topic distribution graph also, we can conclude that LDA is better algorithm for topic distribution than LSA.

From our above analysis, LDA was found to be better for topic modeling so we will use LDA for identifying topics from the given data of news headlines. The entire news headlines would be classified into identified number of topics. Then the Bi-LSTM would be used to train the model in order to make predictions about the topic of particular news headlines.

Though we can have a desired number of topics, but it is necessary to have a balance between number of topics and accuracy of the model. The accuracy of model would be higher for lesser number of topics but as the number of topics are increased, the accuracy of model decreases because the homogeneity of topics distribution also decreases as shown in the figure 4.20. Thus, we have to make a tradeoff between the number of topics and the accuracy of the model. The number of topics should be chosen in such a way that it gives proper insight about the data.

Number of Topics (using LDA)	Accuracy (using Bi-LSTM) (%)
4	92.94
8	87.74
12	87.32
16	85.31
20	82.21
24	81.34
28	80.32
32	78.61
36	75.43
40	74.01

Table 4.1 Accuracy for varying number of topics

The table 4.1 shows that the accuracy of Bi-LSTM model decreases with increment in number of topics. The topics 24 with accuracy 81.34% looks good for classification as the accuracy of model is quite good and also 24 topics shows clear insight into the topics of the data. The separation of topics is also good for 24 topics. Beyond 24 the topics start mixing up with each other, which makes it difficult to distinguish among the topics.

4.3.5 Result of Bi-LSTM model for Topic Modeling

After going through the clustering and bar graphs of both LDA and LSA, it was found that LDA model is performing better than LSA model for any number of topics. So, LDA model was used for classification of news headlines dataset into a number of identified topics. In LDA as well after striking balance between number of topics and accuracy, the model with 24 topics was found to be better as the topics are more easily identifiable. Thus, the entire dataset was classified into 24 topics. Once it was done, the newly classified dataset was trained using Bi-LSTM model so that unseen news headlines could be categorized into one of 24 topics.

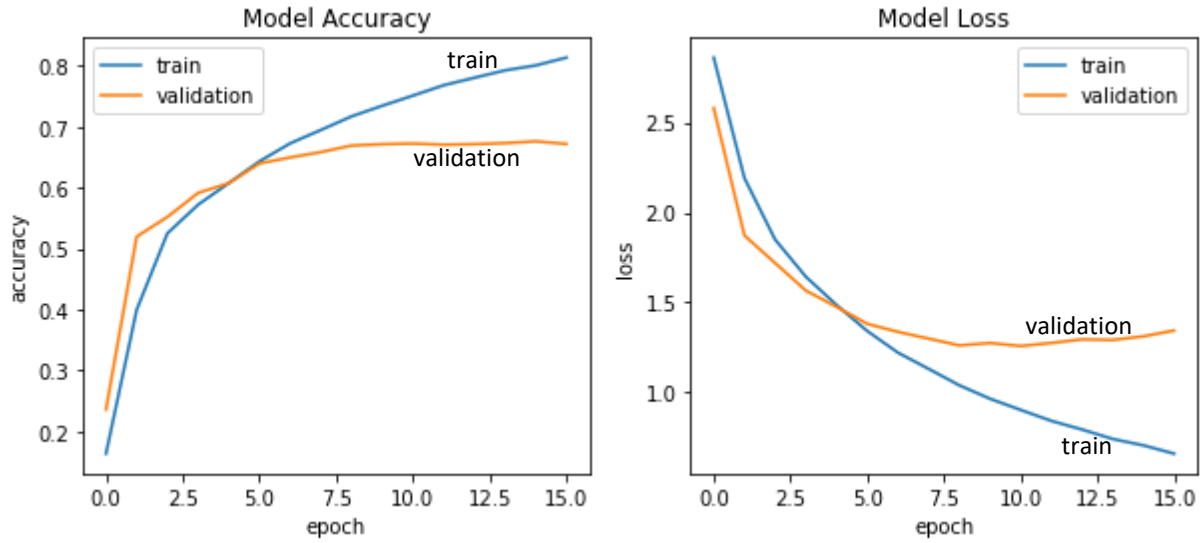


Fig 4.21 Model accuracy and loss graph for topic modeling

The callback function was used so that the training of model stops once it's accuracy or loss value is satisfied.

No. of epochs = 16
Batch size = 64
Learning rate= 0.1

Precision = 89.25%
Recall = 74.17%
Model Accuracy = 81.34%

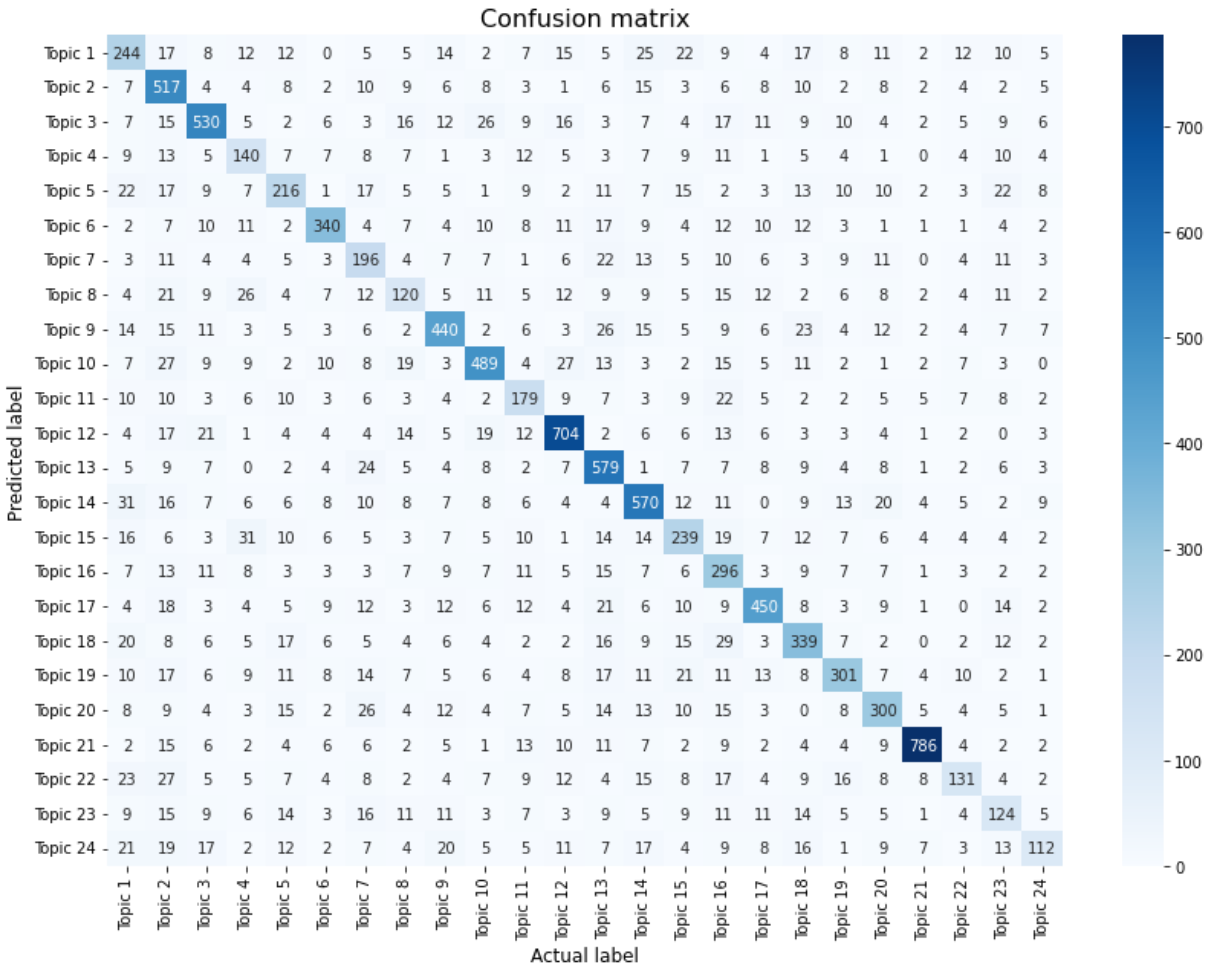


Fig 4.22 Confusion matrix for 24 topics (LDA)

When the bar graphs for varying number of topics are observed, it can be seen that the graph is more uniformly distributed when the number of topics is less. As the number of topics are increased the graph become non-uniform. Also, the accuracy of model decreases as the number of topics are increased, so it is necessary to maintain balance between number of topics and accuracy of the model. Both these factors cannot be considered into isolation, they should be evaluated together. If we want more precise classification of news headlines, we may have to compromise on accuracy to some extent and vice-versa. Thus, after taking various factors into consideration, twenty-four topics was found to be better classification with good accuracy of model as well.

The major words symbolizing each of 24 topics have been listed below:

TOPICS	WORDS
Topic 1	india, nagpur
Topic 2	style, life
Topic 3	hyderabad, india
Topic 4	visakhapatnam, gurgaon
Topic 5	world, cricket
Topic 6	india, ghaziabad
Topic 7	entertainment, movies
Topic 8	meerut, bareilly
Topic 9	crore, worth
Topic 10	citizen, stories
Topic 11	delhi, chennai
Topic 12	sports, india
Topic 13	business, india
Topic 14	woman, dies
Topic 15	news, tv
Topic 16	punjab, haryana
Topic 17	education, mumbai
Topic 18	jaipur, lucknow
Topic 19	bollywood, hindi
Topic 20	rain, kerala
Topic 21	patna, bihar
Topic 22	kolkata, india
Topic 23	ahmedabad, gujarat
Topic 24	dehradun, uttarakhand

Table 4.2 Topics with their corresponding words

4.4 Examples of News headlines prediction

1. The bomb on plane threat creates flutter at Bengaluru airport
   - Sentiment prediction: Negative
   - Topic prediction: Topic 10

2. India to see major growth in manufacturing sector in the year
   - Sentiment prediction: Positive
   - Topic prediction: Topic 3

3. ISROs SSLV missile launch failed
   - Sentiment prediction: Negative
   - Topic prediction: Topic 15

Chapter 5: CONCLUSION AND FUTURE WORK

5.1 Conclusion

The comparative analysis of supervised and unsupervised learnings was performed for text analysis. For supervised learning, sentiment analysis was performed on labeled data of news headlines and it was found that Bi-LSTM model outperformed the machine learning ensemble model with accuracy 84.92% against 81.67% accuracy of ensemble model. The reason for Bi-LSTM model outperforming ensemble model is that the Bi-LSTM model takes the context of occurrence of words, both before and after words, into consideration.

Similarly, for unsupervised learning, two algorithms were used for topic modeling i.e., LSA and LDA and it was found that LDA performed better and produced more homogeneous and balanced distribution of topics in comparison to LSA. The topics were more distinctly visible in case of LDA than LSA as seen through clustering graphs for both the algorithms for varying number of topics. Also, in case of LDA it was necessary to make a tradeoff between accuracy and number of topics for better topic classification and considering this, 24 topics with accuracy 81.34% was found to be a better choice because twenty-four topics provided clear insight into the dataset and also each topic were distinctly visible. Though, twenty-four topics were found to be suitable for this dataset, it is not always true for other datasets. The number of topics to be selected depends upon the varying number of issues or subjects that have been discussed within that dataset. This research shows the way and method to decide upon number of topics for a particular dataset rather than performing hit and trial method.

5.2 Future Work

Though various efforts have been made to provide a comparative analysis of supervised and unsupervised learning for text analysis through this thesis work, there are still some areas where we can improve and work in future. We can work on multilingual dataset to analyze how the model accuracy changes. We can perform aspect-based sentiment analysis for better understanding of sentiments of text. Also, we can perform multi-class classification to get more minute analysis of sentiment analysis.

We can study the clustering of LSA for varying size of datasets to see whether it performs better with small or large datasets. Also, the number of topics can be increased beyond forty to see the extent of mixing up of topics as the number of topics increases. Along with these, we can perform topic modeling on datasets of various news portals and compare their results to see whether the number of topics can be generalized for various news portals.

REFERENCES

[1] M.E. Sunil, S. Vinay, S, "Kannada Sentiment Analysis using vectorization and Machine Learning", Advances in Intelligent Systems and Computing, vol. 1408, 2021

[2] G. Acharya, "Opinion mining for the detection of hate speech over social media", Thesis, IOE, Pulchowk Campus, 2019

[3] A. Go, R. Bhayani, L. Huang, "Twitter Sentiment Classification using Distant Supervision ", Stanford University

[4] B. Gaye, D. Zhang, A. Wulamu, "A Tweet Sentiment Classification Approach Using a Hybrid Stacked Ensemble Technique", MDPI, Basel Switzerland, Sep, 2021, https://doi.org/10.3390/info12090374

[5] R.N. Behera, M. Roy, S. Dash, "A Novel Machine Learning Approach for Classification of Emotion and Polarity in Sentiment140 Dataset", Conference paper, Nov, 2015

[6] M. Birjali, A. Beni-Hssane, M. Erritale, "Machine learning and semantic analysis-based algorithms for suicide sentiment prediction in social networks", The 8th international conference on emerging ubiquitous systems and pervasive networks, 113 (2017) 65-72

[7] Z. Xu, V.P. Rosas, R. Mihalcea, "Inferring social media users' mental health status from multimodal information", Proceedings of the 12th conference on language resources and evaluation, 6292-6299, 2020

[8] S.T. Rabani, Q.R. Khan, A.M.U.D. Khanday, "Detection of suicidal ideation on twitter using machine learning and ensemble approaches", Baghdad science journal, 17(4):1328-1339, 2020, doi: http://dx.doi.org/10.21123/bsj.2020.17.4.1328

[9] A.M. Schoene, G. Lacey, A.P. Turner, N. Dethlefs, "Dilated LSTM with attention for classification of suicide notes", Proceedings of the 10th international workshop on health text mining and information analysis, 136-145, 2019, doi: https://doi.org/10.18653/v1/D19-62

[10] A.C. Fernandes, R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, D. Chandran, "Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing", Scientific reports, 2018, doi: 10.1038/s41598-018-25773-

[11] M. Taboada, J. Brooke, M. Tofiloski, K. V. M. Stede," Lexicon-Based Methods for Sentiment Analysis", 1 Association for Computational Linguistics,2011

[12] J. Kamps, M. Marx, R.J. Mokken, M. de Rijke," Using WordNet to Measure Semantic Orientations of Adjectives", Language & Inference Technology Group, University of Amsterdam, 2001

[13] G. Xu, Y. Meng, X. Qiu, Z. Yu, X. Wu, “Sentiment analysis of comments text based on BiLSTM”, IEEE access, vol. 7, pp. 51522-51532, 2019

[14] U. Chauhan, A. Shah, “Topic Modeling using Latent Dirichlet Allocation: A survey”, ACM Computing surveys, vol. 54, issue 7, Sep, 2021

[15] H. Jelodar, Y. Wang, “Latent Dirichlet Allocation (LDA) and Topic Modeling: models, applications”, Nov, 2017

[16] I. Vayansky, S.A.P. Kumar, “A review to topic modeling methods”, Information Systems, vol. 94, Dec, 2020