ABSTRACT:

A Blog is an online journal or informational site displaying useful information. It is a platform where a writer shares his views on an individual subject. There are news blogs, story blogs and many other blogs. Twitter is a social media platform that has become increasingly popular with academics as well as students, policymakers, politicians and the general public. Both news articles (news blogs) and twitter provide us with useful information. But it is not always the case that the news articles and tweets of one's liking can be found easily. Our project aims to introduce a recommendation system where users can input their topic of interest and get tweets and news articles of their topic of interest as output. The tweet and news articles recommendation are achieved by machine learning techniques. Generally, a tweet or a news article has no fixed topics and it is mostly based upon the user's mood swings. Also, a user may not stay interested in a particular topic for so long. In this paper, we introduce **LATENT DIRICHLET ALLOCATION (LDA)** for blog and tweet topic model construction. We also use voting algorithm to find similar topics that are liked by similar users. The tweets are also classified as communal and non-communal that can be used to classify tweets during violence or disaster times.

INTRODUCTION:

Newspaper was the main source by which information and news across the world was conveyed for many years. Today, in this Digital world where everyone has access to internet nobody likes to read newspaper. There are many news articles apps where one can read what's happening around the world. Twitter is one of the social media platforms where one can gain lots of information. There are people who sometimes might have the time to read news blog but sometimes prefer tweets with news. So, we devised a new idea of creating a place where the users can get both news blog and tweets recommendation in the same place. We used LDA algorithm to classify the tweets and news articles into similar topics and then recommend them based on their topic name that we assign to them. We tune the recommendation system by issuing more recommendations for the users using voting algorithm. We also use communal and non-communal classification to create a classifier system.

PROBLEM STATEMENT:

Implementation of a social network system where the users are recommended tweets and news blogs according to their interests and after a period of time recommendation based on their likes and dislikes with classification of communal and non-communal tweets.

LITERATURE SURVEY:

LDA topic model for blog recommendation and Countering Communal Microblogs During Disaster Events [BP - 1]

The LDA topic model is a kind of Bayesian model. It is composed of three levels, such as documents, topics and words. A document consists of multiple topics. A topic consists of multiple words. we analyse the probability that a word belongs to a particular topic and the probability that a topic belongs to a particular tweet/news blog. We then cluster the topics using k means and then use indirect recommendation to recommend blogs/tweets to user. The huge number of tweets posted during a disaster event includes information about the present situation as well as the emotions/opinions of the masses. This paper focuses on such category of tweets, which is in sharp contrast to most of the prior research concentrating on extracting situational information. Considering the potentially adverse effects of communal tweets during disasters, in this paper, we develop a classifier to distinguish communal tweets from noncommunal ones, which performs significantly better than existing approaches. We believe that such a system is really helpful for government and local monitoring agencies to take appropriate decisions like filtering or promoting some particular contents.

Pros:

LDA is better than any other topic modelling algorithm. It uses indirect recommendation which is more efficient than direct recommendation. It does not allot a single topic domain to a blog/tweet. A potential way of countering communal content would be to utilize such anti-communal content. This paper helps in promoting such anti-communal content.

Collaborative Filtering [BP - 2]

The Collaborative Filtering is the most successful algorithm in the recommender systems' field. But it suffers from its poor accuracy and scalability. This paper considers the users are m (m is the number of users) points in n dimensional space (n is the number of items) and represents an approach based on user clustering to produce a recommendation for active user by a new method. It uses k-means clustering algorithm to categorize users based on their interests. Then it uses a new method called voting algorithm to develop a recommendation.

Pros:

The Collaborative Filtering is the most successful algorithm in the recommender systems' field. A recommender system is an intelligent system can help users to come across

interesting items. It uses data mining and information filtering techniques. The collaborative filtering creates suggestions for users based on their neighbours' preferences.

Social Media Analysis and Classification of communal and non-communal tweets [BP – 3, RP-1]

This paper is oriented towards analysing social media in order to allow users to create their own preferences to follow (analyse) a specific social media source. The web application has been developed to allow a user to follow specific twitter accounts and categorize the tweets on those accounts based on the user defined taxonomies. The benefit of this project is that any user can track in real time when people are talking about some topic, and it enables anyone to have better insight about society as a whole, their values, norms, what they find interesting, and many other things. The second paper aims to classify tweets gathered during disaster events and crises into communal and non-communal tweets and screen the communal tweets from the user feeds so as to avoid communal riots.

Pros:

This paper introduces the social media analysis web application which analyses tweets according to categories which are created by each user. If categories are defined well, and target group and associated twitter accounts are selected carefully, very valuable information can be retrieved from reports. This application will enable any ordinary user to analyse social media activities. It will help individuals as well as organizations to follow certain events and be aware of others' opinion.

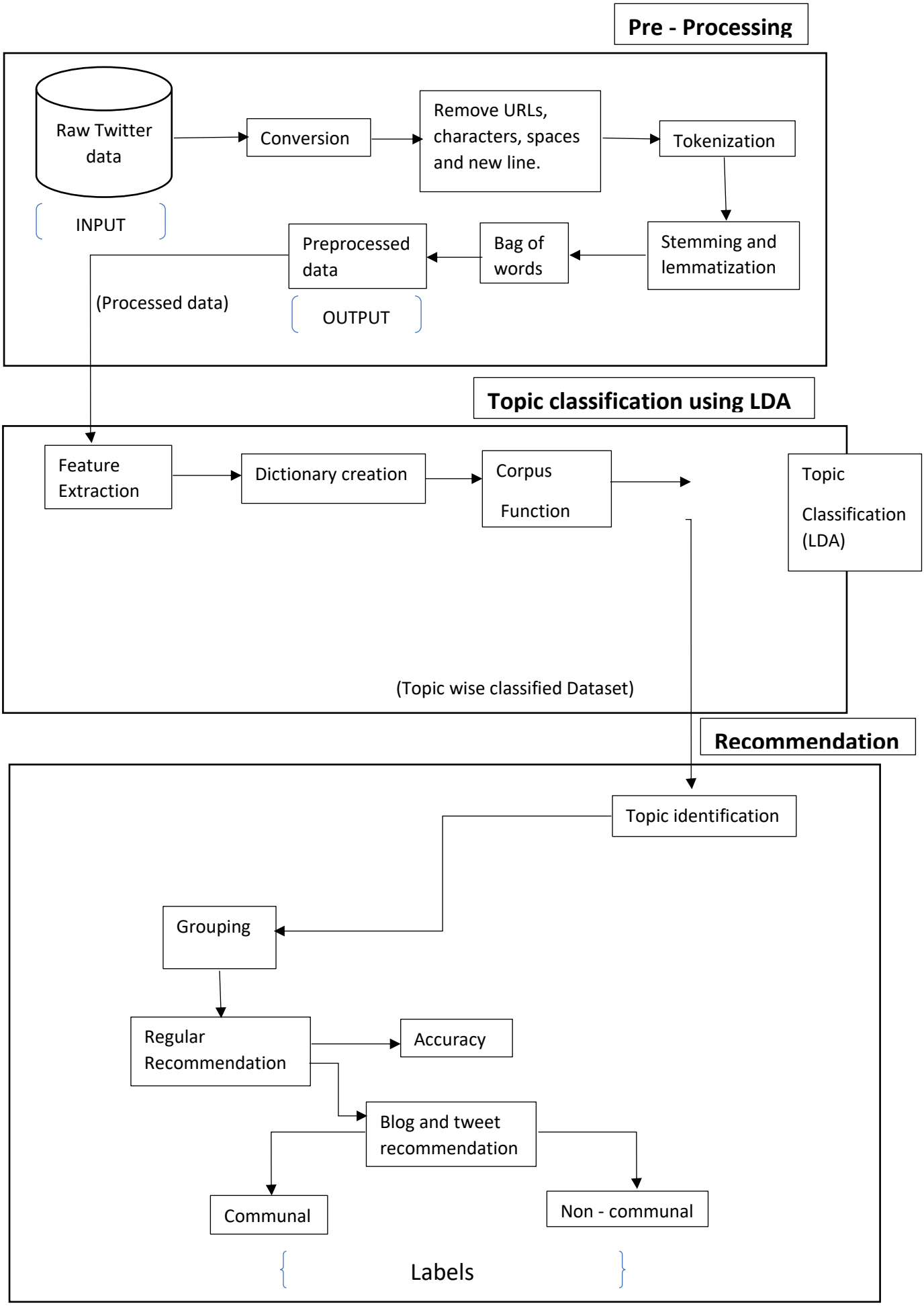
ISSUES IDENTIFIED:

The LDA model needs more calculation and it involves an extra clustering step. It cannot recommend new topics that the user might like. The collaborative filtering algorithm has scalability and sparsity problem. The likes of users have to be stored which may become large if large number of users are present. The tweets found during normal times does not have that many communal tweets. The amount of stored data will increase proportionally with the increase in a number of users. Tweets are known to be informally written and noisy in nature, containing misspellings, abbreviations etc.

SUMMARY:

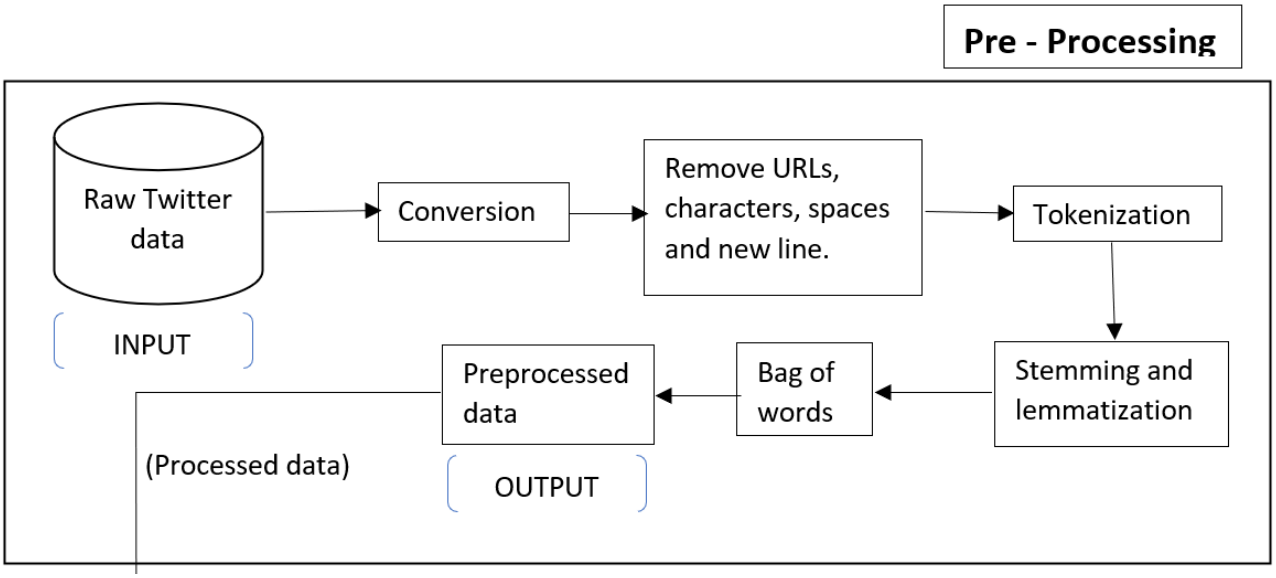
We manually check for the similar groups and reduce the number of topics in LDA and avoid the extra clustering step. We use voting algorithm to provide an extra recommendation in the backend to fine tune our recommendation system. To develop a system for classifying communal and non-communal tweets, we used a separate twitter dataset to improve accuracy of our classifier.

BLOCK DIAGRAM:



MODULES WITH INPUT AND OUTPUT (PSEUDO CODE):

I. PRE-PROCESSING:



- In this module we obtain the raw twitter data and blog data (i.e., input) and convert them into pre-processed output for the next module.
- In-shorts is a popular news blog platform and we take in-shorts dataset.
- We develop our own tweet dataset.
- This conversion is done by performing some operations such as removing URLs from the tweet, removing unnecessary characters, spaces and lines.
- Then we tokenize the tweet and perform stemming (a process where words are reduced to a root through dropping unnecessary characters, usually a suffix) and lemmatization (process of converting a word to its base form by considering the context and converting the word to its meaningful base form).
- Finally, we get a bag of words as the pre-processed output which we will use as input for the next module.

PSEUDOCODE:

1.Import the dataset.

2.Create data-frames and convert all letters to lower case.

3.Remove punctuation using RE.

4.Tokenize tweets using Tweet tokenizer.

5.Lemmatization using NLTK library.

```
lemmatizer = nltk.stem.WordNetLemmatizer()

w_tokenizer = TweetTokenizer()

def lemmatize_text(text):

 return [(lemmatizer.lemmatize(w)) for w in \
```

w_tokenizer.tokenize((text))]

6.Remove stop words.

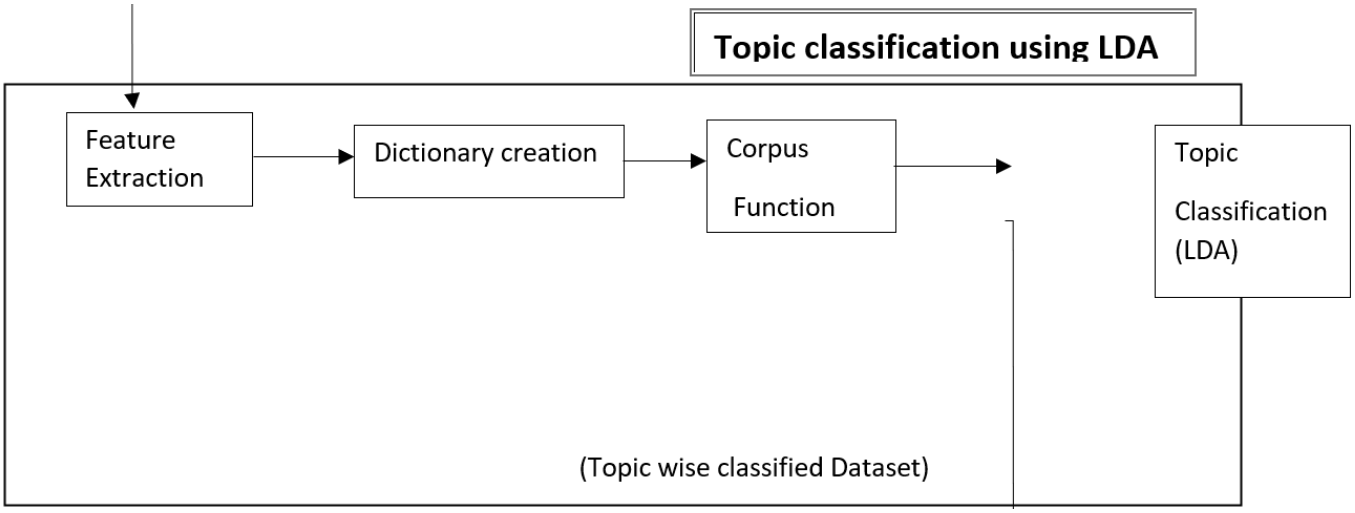
Input -> Raw Twitter Data and Blog Posts.

Output -> Pre-Processed Data

Contribution and Innovation:

We are creating our own dataset for the tweet recognition, classification and recommendation. We got a twitter developer account to get the credentials to make our app contact the twitter api and get the required tweets.Then we created a mysql database and the stored the tweets we got from prominent news sources and other important organizations. Since we need not raw tweets / articles but processed data we start the preprocessing by removing URLs, unnecessary characters, stop words(is, was, on, the, etc.).Before doing the above step, we have converted the entire tweet into lowercase since case plays little to no significant part in analysis. After removing unnecessary stuff, we do stemming and lemmatization to find base words of all the words. Finally, after we have tokenized, we have the required input for next module. Tokens of the tweet. Then we select the required features using various models.

II. TOPIC CLASSIFICATION USING LDA:



- Here, we get the pre-processed data as the input, of which we will remove unnecessary features and select the features that are needed.
- We will be using the data with the selected few features for LDA.
- We use the dictionary and the corpus function to group bag of words.
- We then apply LDA to the training set to obtain the different topics with their bag of words.

- These words are the frequent words and their probabilities are also displayed.

PSEUDOCODE:

1.Printing common words in articles as well as in tweets.

2.Create a dictionary

3.Filter dictionary and apply BOW_CORPUS function.

4.Apply LDA model

lda_model = gensim.models.LdaMulticore(bow_corpus_train, num_topics = 6, id2word = dictionary, passes = 10, workers = 2)

5.Topic mapping

6.Test BOW_CORPUS function with testing set.

```
document_num = 4

bow_doc_y = bow_corpus_test[document_num]

for i in range(len(bow_doc_y)):

   print("Word {} (\"{}\") appears {} time.".format(bow_doc_y[i][0],

                                                    dictionary[bow_doc_y[i][0]],

                                                    bow_doc_y[i][1]))
```

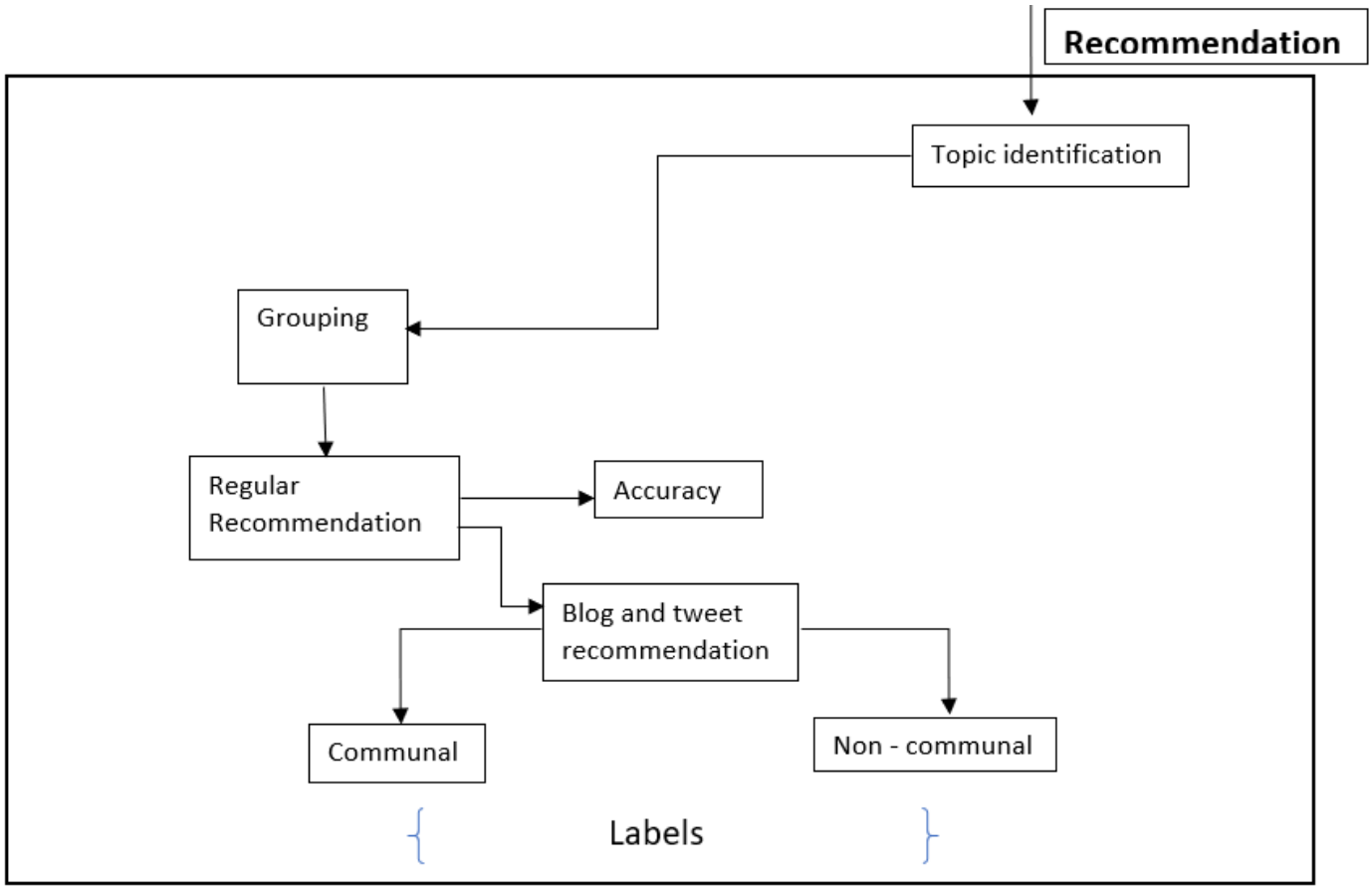
Input-> pre-processed data

Output -> classified topics with frequent words

Contribution and Innovation:

Then using the tokens, we generated we perform LDA. We perform the topic analysis using LDA and store the topics in the database. We create a place where **both news blogs and tweets** can be viewed according to user interests. We recommend these blogs and tweets to users using new and more efficient ML algorithms. Using effective method (LDA) **Latent Dirichlet Allocation** for finding the topic of the blog post / tweet instead of tradition algorithm. In LDA, we use **general topics** so that we can **avoid an extra clustering step** to cluster similar topics.

III. RECOMMENDATION:



- We first identify the topics and then we find the common words and assign a probability to them.
- Then we cluster the posts / tweets together by topics using K-means algorithm and recommend them to users.
- We recommend both blog posts and tweets to users based on their interests (which they'll register during signing up) through indirect recommendation which the topics visited and their frequency is used to find the topics of the post/tweet.
- We also mention if the tweet recommended is communal or non-communal on the side.
- Finally, we check for the accuracy of our algorithms.

PSEUDOCODE:

1. Identify name for topics
2. Upload result of LDA into database
3. Get interest from users
4. Train recommendation system
5. Display recommendation on the front end
6. Classify communal, non-communal

7. Perform Voting algorithm and get new recommendations

```
def get_news_recommendations(postid, num, indices):
    indx = indices[postid]
    sim_scores = list(enumerate(cosine_similarities_news[indx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:num+1]
    print(sim_scores)
    indices = [i[0] for i in sim_scores]
    print("Recommending " + str(num) + " posts similar to \"" + news_item(postid) + "\" ...")
    return news.iloc[indices]
```

Input-> Classified topics with frequent words.

Output-> Recommendation for users

CONTRIBUTION AND INNOVATION:

We allot topic names for each topic in the lda model. We also group each news article and tweet to a particular topic. With the topics we have stored in the database we recommend the customized articles to each user. Using voting algorithm based on our created dataset that contains likes of users for various posts, we perform another recommendation of posts. We use a new dataset that contains more communal tweets to create a tweets classifier that classifies tweets as communal and non-communal. This classifier can be used on our project during times of disaster or communal violence to screen the tweets.

INTERMEDIATE RESULTS:

MODULE-1: PRE-PROCESSING

We use in shorts dataset for news articles and use the newly created dataset using tweets extracted using twitter developer account.

We use java in spring framework to extract tweets and store them in MySQL:

```
2021-03-23 15:02:14.034  INFO 3380 --- [           main] o.s.b.w.embedded.tomcat.TomcatWebServer  : Tomcat started on port(s): 8080 (http) with context path ''
2021-03-23 15:02:14.041  INFO 3380 --- [           main] c.e.i.InteractivearcApplication          : Started InteractivearcApplication in 3.094 seconds (JVM running
2021-03-23 15:02:20.197  INFO 3380 --- [nio-8080-exec-1] o.a.c.c.C.[Tomcat].[localhost].[/]       : Initializing Spring DispatcherServlet 'dispatcherServlet'
2021-03-23 15:02:20.197  INFO 3380 --- [nio-8080-exec-1] o.s.web.servlet.DispatcherServlet        : Initializing Servlet 'dispatcherServlet'
2021-03-23 15:02:20.198  INFO 3380 --- [nio-8080-exec-1] o.s.web.servlet.DispatcherServlet        : Completed initialization in 1 ms
Next Token -> b26v89c19zqg8o3fosqsp085aujrgs491nf9ha7bdpda5
b26v89c19zqg8o3fosqsp085aujrgs491nf9ha7bdpda5
Next Token -> b26v89c19zqg8o3fosqsp07uieyrm997hi6ggxy5jouwt
Next Token -> b26v89c19zqg8o3fosqsp0799p71l2nu1tgrs042id56l
Next Token -> b26v89c19zqg8o3fosqsp06yh9lqyllfbandeu5m8d0u5
Next Token -> b26v89c19zqg8o3fosqsp06nltu8ujsz4fjk3jptu0ltp
Next Token -> b26v89c19zqg8o3fosqsp06cqd06one7q5gg4wjqk3wql
Next Token -> b26v89c19zqg8o3fosqsp061tem38r6engp66bhw77p19
Next Token -> b26v89c19zqg8o3fosqsoy3q5iftwb4920i91po6d92f1
Next Token -> b26v89c19zqg8o3fosqsoy34hni9uy174lnjrnvlvsvi5
Next Token -> b26v89c19zqg8o3fosqsoy1xdfcneb0gjg0qvacz8vmgt
Finished Successfully
```

```
mysql -uroot -pdhaya -h127.0.0.1 -P5000
mysql> select * from test limit 4\G
*************************** 1. row ***************************
          id: 26252
   author_id: 39743812
  like_count: 0
        link: https://twitter.com/39743812/status/1374293279513088000
 quote_count: 0
 reply_count: 0
retweet_count: 0
  tweet_body: The doubling time of COVID-19 cases in India has decreased from 504.4 d

https://t.co/1XJgIQ7gvB
    tweet_id: 1374293279513088000
  tweeted_at: 2021-03-23T09:33:45.000Z
```

We initially pre-process the dataset converting to lower case, removing punctuation, stop words, hashtags (if any) for both datasets (news articles and tweets). The final result of this module is as follows:

News Articles:

	article_tokens	article_final_tokens	headline_tokens	headline_final_tokens
0	[deepminds, ai, system, alphafold, has, been, ...	[deepminds, ai, system, alphafold, ha, recogni...	[50, -, year-old, problem, of, biology, solved...	[50, -, year-old, problem, biology, solved, ar...
1	[microsoft, teams, will, stop, working, on, in...	[microsoft, team, stop, working, internet, exp...	[microsoft, teams, to, stop, working, on, inte...	[microsoft, team, stop, working, internet, exp...
2	[china, in, response, to, reports, of, us, add...	[china, response, report, u, adding, chinese, ...	[hope, us, wont, erect, barriers, to, cooperat...	[hope, u, wont, erect, barrier, cooperation, :...
3	[the, global, smartphone, sales, in, the, thir...	[global, smartphone, sale, third, quarter, 202...	[global, smartphone, sales, in, q3, falls, 57,...	[global, smartphone, sale, q3, fall, 57, %, 36...
4	[the, european, union, (, eu,), is, hoping, t...	[european, union, (, eu,), hoping, u, preside...	[eu, hoping, biden, will, clarify, us, positio...	[eu, hoping, biden, clarify, u, position, digi...
5	[the, members, of, the, joint, parliamentary, ...	[member, joint, parliamentary, committee, divi...	[parliamentary, panel, divided, over, key, iss...	[parliamentary, panel, divided, key, issue, da...

Twitter dataset (Pre-processed data):

	Content	Tokens	hashtags
0	opinion: its 15 years since twitter launched ...	[opinion, its, 15, years, since, twitter, laun...	[]
1	mideast stocks: major gulf markets ease in ear...	[mideast, stocks, major, gulf, markets, ease, ...	[]
2	new york has reported its first confirmed case...	[new, york, has, reported, its, first, confirm...	[]
3	icymi: 'beyond walls' artist saype spray-paint...	[icymi, beyond, walls, artist, saype, spray-pa...	[]
4	saudi aramco reports 49bn profit slump in 2020	[saudi, aramco, reports, 49bn, profit, slump, ...	[]
5	outdoor shows and decoy audiences herald retur...	[outdoor, shows, and, decoy, audiences, herald...	[]

Module -2: Topic Classification Using LDA:

We create a dictionary and find the common words. We then filter the dictionary and apply bow corpus function. We then do LDA classification. This process is repeated for both tweets and news articles. We manually identify the number of topics to use for lda model. We do this by trial and error.

NEWS ARTICLES: 8 topics was found to be ideal

```
Topic: 0
Words: 0.017*"$" + 0.017*"farmer" + 0.014*"government" + 0.014*"delhi" +
0.014*"billion" + 0.013*"minister" + 0.010*"congress" + 0.009*"cm" + 0.009*"state" +
0.008*"law"


Topic: 1
Words: 0.018*"film" + 0.010*"bjp" + 0.009*"world" + 0.008*"party" + 0.007*"time" +
0.007*"actress" + 0.007*"president" + 0.007*"leader" + 0.007*"day" + 0.006*"actor"


Topic: 2
Words: 0.014*"vaccine" + 0.013*"company" + 0.010*"19" + 0.010*"covid" + 0.009*"police"
+ 0.009*"country" + 0.009*"use" + 0.009*"vehicle" + 0.007*"world" + 0.007*"last"

Topic: 3
Words: 0.011*"company" + 0.009*"user" + 0.009*"apple" + 0.008*"facebook" +
0.008*"government" + 0.008*"google" + 0.007*"app" + 0.007*"show" + 0.007*"country" +
0.006*"uk"

Topic: 4
Words: 0.033*"vaccine" + 0.019*"19" + 0.018*"covid" + 0.013*"actor" +
0.010*"coronavirus" + 0.007*"uk" + 0.006*"health" + 0.006*"group" + 0.006*"would" +
0.006*"positive"

Topic: 5
Words: 0.019*"trump" + 0.018*"election" + 0.014*"video" + 0.012*"picture" +
0.011*"took" + 0.011*"president" + 0.010*"twitter" + 0.010*"wrote" + 0.009*"donald" +
0.008*"one"


Topic: 6
Words: 0.010*"film" + 0.008*"coronavirus" + 0.008*"death" + 0.007*"tweeted" +
0.007*"minister" + 0.007*"singh" + 0.007*"actor" + 0.007*"space" + 0.006*"2" +
0.006*"%"
```

```
Topic: 7
Words: 0.015*"%" + 0.012*"china" + 0.011*"₹" + 0.011*"official" + 0.010*"crore" +
0.009*"time" + 0.008*"shared" + 0.007*"facebook" + 0.007*"instagram" + 0.007*"picture"
```

TWEETS: 6 topics was found to be ideal

```
Topic: 0
Words: 0.024*"india" + 0.023*"read" + 0.017*"indias" + 0.015*"story" + 0.013*"police" +
0.008*"mandate" + 0.008*"said" + 0.008*"england" + 0.007*"calling" + 0.007*"ensure"


Topic: 1
Words: 0.022*"vaccine" + 0.008*"astrazeneca" + 0.008*"report" + 0.008*"hours" +
0.008*"use" + 0.008*"us" + 0.008*"says" + 0.007*"million" + 0.007*"covid" +
0.007*"according"


Topic: 2
Words: 0.027*"cases" + 0.019*"case" + 0.019*"details" + 0.015*"maharashtra" +
0.013*"#covid19" + 0.010*"rs" + 0.009*"covid" + 0.009*"kumar" + 0.008*"#coronavirus" +
0.008*"scare"


Topic: 3
Words: 0.062*"election" + 0.043*"#may2withtimesnow" + 0.042*"times" + 0.042*"amp" +
0.034*"news" + 0.025*"polls" + 0.019*"india's" + 0.018*"1" + 0.017*"pulse" +
0.016*"ground"


Topic: 4
Words: 0.024*"bjp" + 0.024*"minister" + 0.019*"bengal" + 0.016*"said" + 0.013*"west" +
0.012*"party" + 0.012*"tmc" + 0.012*"congress" + 0.012*"government" + 0.011*"chief"


Topic: 5
Words: 0.024*"pm" + 0.020*"india" + 0.017*"rahul" + 0.013*"join" + 0.013*"shivshankar"
+ 0.010*"special" + 0.010*"watch" + 0.009*"upfront" + 0.008*"pradesh" + 0.008*"schools"
```

Module-3: Recommendation:

Topic identification: we identify the topics for each category in the generated lda model and map them with names and store them in the MySQL database. This process was repeated for both tweets and news articles.

News articles: the topic is added to the content.

In [29]:
```
# test = pd.DataFrame()
final_topics['headline']=test_df.iloc[:,1:2]
final_topics['article']=test_df.iloc[:,2:3]
final_topics.head()
```

Out[29]:

	topic	headline	article
0	Technology	Porn star jokes she could post nudes on Insta ...	Porn star Kendra Sunderland's Instagram accoun...
1	Buisness	Satellite pics show remains of SpaceX's rocket...	Satellite images showed the remains of Elon Mu...
2	Politics	Not in favour of recommending such games: Top ...	National Commission for Protection of Child Ri...
3	Covid 19	Researchers develop first AI tool to detect CO...	A team of researchers from the University of V...
4	Buisness	Virgin Galactic's test space flight cut short ...	Richard Branson's space tourism company Virgin...

Tweets: the topic name is found and added to the content

Out[84]:

	topic	content
0	World News	there are also events being organised where re...
1	India	the uae had in 2016 created the post of minist...
2	India	aiadmk releases manifesto for upcoming #tamiln...
3	Covid 19	#punjab chief secretary vini mahajan on sunday...
4	Covid 19	rt : #indvseng\r\n\r\nwhen ishan kishan burst ...

First recommendation:

A user login/signup form is created and the topic names are added so that the users can select their choice of interests. we then use the database to access the news and tweets and display them in the simple frontend. We check with a supplied topic name in the backend first.

```
inp=input('Enter interested domain: ')
recommend(inp,data)
```

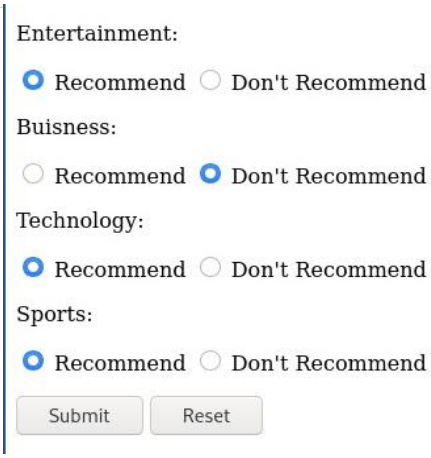
```
Enter interested domain: Sports
 0 Dharmendra, Who Turned 85 Years Old On December 8, Has Said His Daughter Ahana Deol'S Newborn Twins, Astraia And Adea, Are The Biggest Gifts For
 1 The Makers Of Shahid Kapoor'S Upcoming Film 'Jersey' Deferred A Shooting Schedule In Chandigarh Amid Farmers' Protest And Flew To Dehradun. The
 2 An Australian Woman Was Arrested For Allegedly Using The Dark Web To Hire A Contract Killer To Murder Her Parents For Financial Gains, Following
 3 Archaeologists Have Discovered Well-Preserved Remains Of A Possibly Rich Man And His Slave Scalded To Death By Mount Vesuvius' Eruption Nearly 2
 4 Anushka Sharma Took To Social Media To Share A Throwback Picture Of Herself Doing Shirshasana During Her Pregnancy With The Support Of Her Husba
 5 Talking About The Theatrical Release Of His Upcoming Comedy Film 'Indoo Ki Jawani', Director Abir Sengupta Said He'S Happy That The Producers Ha
 6 As Sharman Joshi'S 2015 Thriller-Romance Film 'Hate Story 3' Completed Five Years, The Actor Recalled Signing The Film And Said, "A Lot Of Peopl
 7 Music Composers And Singers Sachet Tandon And Parampara Thakur, Known For Composing 'Bekhayali' And 'Mere Sohneya' In 'Kabir Singh', Got Married
 8 A 13-Year-Old Palestinian Was Shot And Killed By Israeli Soldiers During A Clash In The West Bank, The Palestinian Health Ministry Said On Frida
 9 After The Pictures Of His Look From His Upcoming Film 'Bob Biswas' Surfaced Online, Abhishek Bachchan Said, "I Would'Ve Loved For...Look To Come
```

FRONTEND:

Getting Username and Password

Received Data:

After pressing next we redirect to the next page where we get the user's interests.



Tweets

Finance

rt : #indvseng when ishan kishan burst into the domestic team as a fresh-faced teenager his senior teammates spun a nickna

rt : #indveng only former australian captain ricky ponting (15440) and south africa legend graeme smith (14878) had achieved t

Vaccine

our herd immunity has not increased to such a level that we may become careless (about covid-19): former sec indian medical association tells swati joshi on special edition #covidsurgealarm

rt : check out the official photos of #mehreenpirzada and #bhavyabishnoi's engagement and their dance video

India

the uae had in 2016 created the post of minister of state for happiness

aiadmk releases manifesto for upcoming #tamilnadu polls ( ) #itvideo

politics

there are also events being organised where residents are invited to get a protest tattoo to raise funds for the civil disobedience movement

rt : the mid-day meal (mdm) scheme which provides a cooked meal to students of primary and upper primary classes benefited

Covid 19

#punjab chief secretary vini mahajan on sunday reviewed the emergency measures to tackle the peak in #covid19 cases and expressed concern over the third peak with 1515 positive cases reported photo: ians (file)

complete lack of adherence to covid appropriate behaviour is the main reason behind the sudden spike in covid cases: member covid task force maharashtra tells swati joshi on special edition #covidsurgealarm

Election

the conspiracy theory has now been busted it is a big setback to mamata banerjee amp; tmc: journalist amp; political analyst tells swati joshi on spl ed #mamataincidentreportexclusive

it is unfortunate that an accident occurred to mamata banerjee but tmc has been blaming us from the beginning: spokesperson west bengal bjp tells swati joshi on spl ed #mamataincidentreportexclusive

News

Technology

Porn star jokes she could post nudes on Insta as she 'performed sex acts on CEO'

Apple shuts Music Memos app, asks users to export files to Voice Memos

Politics

Not in favour of recommending such games: Top child rights body on PUBG relaunch

K'taka govt to provide protection to iPhone plant Wistron after vandalism

Covid 19

Voting Algorithm: we use created dataset which shows the posts that the user liked to display further new recommendation in posts. We use this algorithm to further find new tweets and news articles which similar users have liked (similar to the YouTube recommendation)

If a user shows interest in vaccine field, we use voting algorithm to show similar piece of interesting tweets through voting algorithm

```
[(11, 0.12002190951147645), (12, 0.03612931290944911), (33, 0.026318895616060947)
]
Recommending 3 posts similar to "Vaccine" ...
```

- Out[43]:

	id	topic	tweet	tweet_id	mix
11	11	Covid 19	complete lack of adherence to covid appropriat...	1371131251113926660	complete lack of adherence to covid appropriat...
12	12	Covid 19	india has registered over 25000 cases in the l...	1371130574874677249	india has registered over 25000 cases in the l...
33	33	Election	it has now been proved that the 'eyewitnesses'...	1371126932373303296	it has now been proved that the 'eyewitnesses'...

```
[(5, 1.0), (6, 1.0), (7, 1.0)]
Recommending 3 posts similar to "Technology" ...
```

	id	topic	headline	article	mix
5	5	Technology	Apple shuts Music Memos app, asks users to exp...	Apple has announced that it's discontinuing it...	Apple has announced that it's discontinuing it...
6	6	Technology	28% Indians hide real identities on social med...	Global cybersecurity firm Kaspersky in its rep...	Global cybersecurity firm Kaspersky in its rep...
7	7	Technology	Malware campaign adding extensions in Chrome, ...	Google Chrome, Firefox and other browsers are ...	Google Chrome, Firefox and other browsers are ...

Communal and Non-communal classification: we use new twitter dataset to also include classification of communal and non-communal tweets. This finds application during communal riots to overview the tweets that are being recommended. As the current news articles and twitter dataset does not have many such communal tweets, this classifier is used by us on a different dataset.

Input file:

disasterTweets.csv - LibreOffice Calc

	A
1	text
2	Communal violence in Bhainsa, Telangana. "Stones were pelted on Muslims' houses and some houses and vehicles were set ablaze...
3	Telangana: Section 144 has been imposed in Bhainsa from January 13 to 15, after clash erupted between two groups on January 12. Po...
4	Arsonist sets cars ablaze at dealership https://t.co/gOQvyJbpVI
5	Arsonist sets cars ablaze at dealership https://t.co/0gL7NUCPlb https://t.co/u1CcBhOWh9
6	"Lord Jesus, your love brings freedom and pardon. Fill me with your Holy Spirit and set my heart ablaze with your l... https://t.co/VlTznnPNi8
7	If this child was Chinese, this tweet would have gone viral. Social media would be ablaze. SNL would have made a racist j...
8	Several houses have been set ablaze in Ngemsibaa village, Oku sub division in the North West Region of Cameroon by... https://t.co/99uHGAzxy2
9	Asansol: A BJP office in Salanpur village was set ablaze last night. BJP has alleged that TMC is behind the incident. Police has b...
10	National Security Minister, Kan Dapaah's side chic has set the internet ablaze with her latest powerful video.... https://t.co/rhzOMQVSlj
11	This creature who's soul is no longer clarent but blue ablaze This thing Carrying memories Memories of... https://t.co/tBKSNDrDoX
12	Images showing the havoc caused by the #Cameroon military as they torched houses in #Oku.The shameless military is reported...
13	Social media went bananas after Chuba Hubbard announced Monday evening his plans to return to #okstate. https://t.co/peN...
14	Hausa youths set Area Office of Apapa-Iganmu Local Council Development Area ablaze. Okada Riders stormed the LG area office...
15	Under #MamataBanerjee political violence & vandalism continues to unabated in West Bengal! office in Asanol was...
16	AMEN! Set the whole system ablaze, man. https://t.co/J08xHDcGbD

Output File:

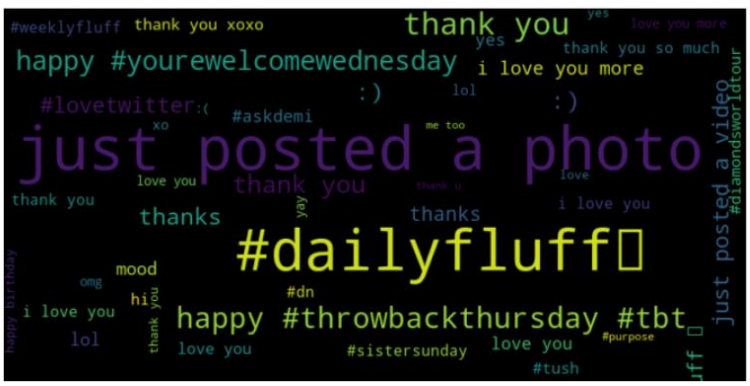
output.txt - LibreOffice Calc

	A	B
1	text	Classification
2	Communal violence in Bhainsa, Telangana. "Stones were pelted on Muslims' houses and some houses and vehicles were set ablaze...	1
3	Telangana: Section 144 has been imposed in Bhainsa from January 13 to 15, after clash erupted between two groups on January 12. Po...	1
4	Arsonist sets cars ablaze at dealership https://t.co/gOQvyJbpVI	1
5	Arsonist sets cars ablaze at dealership https://t.co/0gL7NUCPlb https://t.co/u1CcBhOWh9	1
6	"Lord Jesus, your love brings freedom and pardon. Fill me with your Holy Spirit and set my heart ablaze with your l... https://t.co/VlTznnPNi8	0
7	If this child was Chinese, this tweet would have gone viral. Social media would be ablaze. SNL would have made a racist j...	2
8	Several houses have been set ablaze in Ngemsibaa village, Oku sub division in the North West Region of Cameroon by... https://t.co/99uHGAzxy2	1
9	Asansol: A BJP office in Salanpur village was set ablaze last night. BJP has alleged that TMC is behind the incident. Police has b...	1
10	National Security Minister, Kan Dapaah's side chic has set the internet ablaze with her latest powerful video.... https://t.co/rhzOMQVSlj	2
11	This creature who's soul is no longer clarent but blue ablaze This thing Carrying memories Memories of... https://t.co/tBKSNDrDoX	0
12	Images showing the havoc caused by the #Cameroon military as they torched houses in #Oku.The shameless military is reported...	1
13	Social media went bananas after Chuba Hubbard announced Monday evening his plans to return to #okstate. https://t.co/peN...	0
14	Hausa youths set Area Office of Apapa-Iganmu Local Council Development Area ablaze. Okada Riders stormed the LG area office...	1
15	Under #MamataBanerjee political violence & vandalism continues to unabated in West Bengal! office in Asanol was...	1
16	AMEN! Set the whole system ablaze, man. https://t.co/J08xHDcGbD	0
17	Images showing the havoc caused by the #Cameroon military as they torched houses in #Oku.The shameless military is... https://t.co/gIwZCH533D	1
18	No cows today but our local factory is sadly still ablaze #REDJanuary2020 https://t.co/CMyuKzrcKz	1
19	Rengoku sets my heart ablaze😍❤🔥 P.s. I missed this style of coloring I do so here it is c: #鬼滅の刃 https://t.co/YrUF9g68s0	0

EVALUATION METRICS AND GRAPHS, RESULT TABLE:

1. Word cloud creation:

```
wc = WordCloud(width=800, height=400, max_words=50).generate_from_frequencies(fdist)
plt.figure(figsize=(12,10))
plt.imshow(wc, interpolation="bilinear")
plt.axis("off")
plt.show()
```



2. Word cloud to test headlines:

```
wc = WordCloud(width=900, height=500, max_words=50).generate_from_frequencies(fdist)
plt.figure(figsize=(12,10))
plt.imshow(wc, interpolation="bilinear")
plt.axis("off")
plt.show()
```

CONCLUSION AND FUTURE DIRECTION:

We have completed the project. We started with collecting tweets from twitter and converting into a dataset. We have also then applied lda model to classify tweets to various topics. We gave the topic names and also got interest from user and deployed a recommendation system. We have also tuned a voting algorithm for further recommendation. We have further classified tweets into communal and non-communal. Our next step will be to tune up the project further and use the classifier at times of communal violence or disaster events.

REFERENCES IN IEEE FORMAT:

- Jianyong Duan, Yamin Ai and Xia li, "LDA topic model for microblog recommendation," – Published in 2015 International Conference on Asian Language Processing (IALP), Suzhou, China, 2015, pp. 185-188. (base paper 1)

- G. M. Dakhel and M. Mahdavi, "A new collaborative filtering algorithm using K-means clustering and neighbors' voting," – Published in 2011 11th International Conference on Hybrid Intelligent Systems (HIS), Melacca, Malaysia, 2011, pp. 179-184. (base paper 2)

- B. Isakovic, D. Keco and N. Dogru, "Social media analysis web application," – Published in 2017 XXVI International Conference on Information, Communication and Automation Technologies (ICAT), Sarajevo, Bosnia and Herzegovina, 2017, pp. 1-6. (base paper 3)

- K. Rudra, A. Sharma, N. Ganguly and S. Ghosh, "Characterizing and Countering Communal Microblogs During Disaster Events," – Published in IEEE Transactions on Computational Social Systems(Vol. 5, No. 2, pp. 403-417, June 2018). (reference paper-1)