A Comprehensive Technique for User Activity Based Twitter Content Summarization

Ayushi Gupta, Devyani Keskar, Madhur Firodiya, Siddhi Hagawane

Abstract: *Going through thousands of comments in order to understand opinion of people on a particular post ingests in a lot of time and resources of the user. By developing this system, we aim that user gets updated with summarized information of all such events in a time constrained manner. It involves merging multiple opinions stated on the social platform and summarizing it to provide the gist of the topic in order to improve ergonomic experience. For this purpose, our system displays both abstractive and extractive summary of the content. Extractive summary generation makes use of Page rank algorithm and abstractive summary generation makes use of RNN (LSTM).*

Keywords : *Text analysis, Tweets, Live streaming, Filter based analysis, Anomaly detection*

I. INTRODUCTION

Summarization is the process of generating short, fluent, and most importantly accurate summary of a respectively longer text document . The main idea behind summarization is to be able to find a short subset of the most essential information from the entire set and present it in a human-readable format. As online textual data grows, summarization methods have potential to be very helpful because more useful information can be read in a short time. Social media platforms like twitter, facebook, instagram generate large amount of data through post and comment.

Every post on this platform have comments and scrolling through these comments requires the user to spend a lots of time. [4]If the gist of public opinion on a particular post can be provided to user, he can state his opinion in comparatively short span, thereby enabling him to invest the same amount of time for some other work. Generation of summary is not just limited to comments and captions on posts but also can be used for summarizing large amount of data available on the internet through blogs, Wikipedia and research papers. Taking into consideration a subset here we design a system to generate summary of the tweets and replies on the twitter. [1]We also wish to further expand it for recapitalizing the content on other social media platforms. Such a summarization system can also be used in e-commerce applications for generating the gist of thousands of product reviews which can then be used to analyse the sale of the product and understand the public sentiment about likability of product. It can also help the customer to buy the perfect product. [17]Here we propose a system which is a web application for summarizing tweets on twitter and ensure that they can be summarized effectively and efficiently. In order to achieve these goals, we developed the following objectives:

- Research current technologies and progress associated with tweet summarization
- Perform live streaming to collect dated tweets.
- Implement algorithms and models for different methods of tweet summarization.
- Evaluate the models and tune them if necessary.
- Build and host web application which takes tweets as input and produce summary as output.

II. LITERATURE SURVEY

We carried out the literature survey by going through various research works done in the field of summarization. Some of the drawbacks identified are Linguistic constraint, implementation complexity , underlying limitations of methods applied, semantic & syntactic constraints ,etc. To overcome these shortcomings , we have proposed a system that combines the benefits of multiple models using a “hybrid approach” which implements semantic as well as syntactic approach for summarization.

A. Research Work

[1]Koustav Rudra,Siddhartha Banarjee”Extractive summarization is applied to extract important tweets and then abstractive summarization is applied to improve readability”.

[7]N.Moratanch and S.Chitrakala, "A Survey on Extractive Text Summarization". In this paper, author has described the word level features and sentence level features. In this paper author have categorized all extractive summarization methods into unsupervised and supervised methods and have explained each method and have depicted few evaluation metrics.

Akshil Kumar et al. In this paper author has analyzed and compared the performance of three different algorithms.

Firstly, the different text summarization techniques explained. Extraction based techniques are used to extract important keywords to be included in the summary[12]. For comparison three comparison three keyword extraction algorithms namely TextRank, LexRank, Latent Semantic Analysis (LSA) were used. Three algorithms are explained and implemented in python language. The ROUGE 1 is used to evaluate the effectiveness of the extracted keywords. The results of the algorithms compared with the handwritten summaries and evaluate the performance. In the end, the TextRank Algorithm gives a better result than other two algorithms.

Revised Manuscript Received on April 12, 2020.
* Correspondence Author

Ayushi Gupta*, Student,Dept. Of Computer Engg., MITCOE,Pune Email:ayushimg9@gmail.com
Devyani Keskar, Student,Dept. Of Computer Engg., MITCOE,Pune Email: devyani.keskar@gmail
Madhur Firodiya, Student,Dept. Of Computer Engg., MITCOE,Pune Email: madhurfirodiya15@gmail.com
Siddhi Hagawane, Student,Dept. Of Computer Engg., MITCOE,Pune Email: siddhihagawane22@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Retrieval Number: D7827049420/2020©BEIESP
DOI: 10.35940/ijeat.D7827.049420
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.

[13]Pankaj Gupta et al. In this paper author has reviewed different techniques of Sentiment analysis and different techniques of text summarization. Sentiment analysis is a machine learning approach in which machine learns and analyze the sentiments, emotions present in the text. The machine learning methods like Naive Bayes Classifier and Support Machine Vectors (SVM) are used. These methods are used to determine the emotions and sentiments in the text data like reviews about movies or products. In Text summarization, uses the natural language processing (NPL) and linguistic features of sentences are used for checking the importance of the words and sentences that can be included in the final summary. In this paper, a survey has been done of previous research work related to text summarization and Sentiment analysis, so that new research area can be explored by considering the merits and demerits of the current techniques and strategies.

Thus, considering the mentioned research work in the field in this paper we have proposed a system taking into account the drawbacks and advantages of various approaches,Our proposed system uses LSTM(Long Short Term memory) which is a type of RNN (Recurrent neural networks) that processes information as it proceeds ahead and also considers the previously processed information as feedback to improve the accuracy and precision of results. LSTM is used for generation of abstractive summary. Aiming for better feasibility we chose to use PageRank algorithm for generation of extractive summary.

B. Current Market Survey

- Summary Scanner:

Summary Scanner is a new app that is not only a time-saver but also a game changer. It allows you to automatically summarize any document in seconds without missing out on any vital information. It's described as the most powerful mobile scanner for office, business, and personal use in the world.

- BookBhook:

Bookbhook reading app thoughtfully curated handcrafted book summary of life changing books that take no more than 15 minutes to read, in English and Hindi. Yes, you get to read summary of life changing English books in Hindi for free. You can add book summaries that you enjoyed in the wish list saver feature- add to favorites. Want to get hooked to bite-sized chat story app? bookbhook brings to you free chat stories that you can read.

- Self-Help: Provides Books summaries to readers with concise and comprehensive text..

III. SYSTEM DESCRIPTION

Our system is primarily based on extractive as well as abstractive analysis for which the following system architecture is proposed. It has Majorly the following components/modules:

i) Tweets Extraction Module

Tweets along with its corresponding replies are retrieved by live streaming using the Twitter api for the “userhandle” provided as input by the user.

These are written in an excel file to be used for further processing.

ii) Preprocessing Paradigm

The preprocessing paradigm is used to extract the necessary contents from the collected comments and reduce it in order to make the processing easy.

iii) Extractive Analysis Module

It will generate an extractive summary for a set of tweets and replies to those tweets. The extractive summary will then be used to generate abstractive summary.

iv) Abstractive Analysis Module

The abstractive summary i.e the interpreted summary of the original tweets along with its replies is generated by this module.It is displayed to the user along with the extracted ranked tweets on the system dashboard.

v) Database Module

The excel file storing the monthly tweets along with the summary will be converted into json file and mongodb will be used to import the file.

vi) User Interface

The dashboard will be the platform where user inputs will be recorded to process and deliver the output back to the users on a single click.

The UI will also take in user feedback on every result generated to calculate and improve the accuracy of the system .

Following constraints will be taken into consideration while designing the architecture:

- After every month ,the monthly activity of the requested users till date will be updated
- Summary generated will be limited to max 1000 chars..

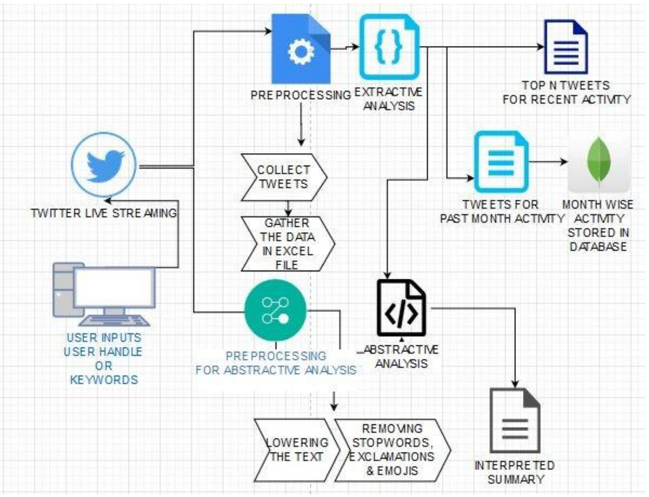

PREPROCESSING
EXTRACTIVE ANALYSIS
TOP N TWEETS FOR RECENT ACTIVITY
COLLECT TWEETS
GATHER THE DATA IN EXCEL FILE
TWITTER LIVE STREAMING
TWEETS FOR PAST MONTH ACTIVITY
MONTH WISE ACTIVITY STORED IN DATABASE
USER INPUTS USER HANDLE OR KEYWORDS
PREPROCESSING FOR ABSTRACTIVE ANALYSIS
ABSTRACTIVE ANALYSIS
LOWERING THE TEXT
REMOVING STOPWORDS, EXCLAMATIONS & EMOJIS
INTERPRETED SUMMARY

Figure 1. System Architecture

IV. METHODOLOGY

We initially extract the tweets from twitter by live streaming using twitter-API and tweepy. For this purpose, we have to configure the twitter app. Once we create an application, twitter provides Consumer Key, Consumer Secret, Access Token and Access Token Secret. They are four data elements that are used to extract tweets from twitter. The data that is extracted is written in an excel file . However, the data which is obtained has some parts which are not required for summarization purpose.

Hence, we need to remove the irrelevant details and preserve the main sections of tweets.The data is summarized using two different methods i.e. Extractive and Abstractive. These methods will be used to process both monthly activity of a specified user and collective data of tweets tweeted and replies given to the user in past seven days i.e.recent data.

Extractive Summary Generation: Produces a subset of the sentences from the original text.Page-Rank algorithm is used for this purpose.

1. J={L1,2,3…….n)
Where n belongs to J.
2. sentence[]={L1,L2,L3...Ln}
3. Sentence[]-{stopwords}
4. Construct word vectors T1={w1,w2,w3…..wn} . . Tn={w1,w2,w3…..wn}
5. Construct similarity matrix using cosine similarity
Cosine_similarity = 1- m.n /|m||n|
Where m= weight of sentence 1 n= weight of sentence n
6. Construct a graph G = {Sn,En}
where Sn=Sentence
En=Similarity Score
7. ranked_sentences[]=[R1,R2,R3...Rn] Where R1>R2>R3…>Rn

The excel file is read from tweets.Sentence Tokenizer will be used for tokenizing the sentences and vectors will be created.Word embeddings are used to convert phrases into numerical formats.thus helping the networks to learn better.It also provides certain characteristics of the words used in vocabulary.We have used GloVe for word representation in our code which is provided by Stanford.We have limited our dimensions to 100.This helped us to assign weights to the sentences.Now based on weights assigned,similarity matrix based on cosine similarity is constructed.Using this matrix sentences are ranked and top 5 sentences are given as output if the number of tweets is greater than 5 else n tweets are given as output if number of tweets is less than 5(n<5).The output is written in the same excel sheet provided as input.

Abstractive Summary Generation: Reproduces important material in a new way after interpretation. Examines text using advanced natural language techniques to generate a new shorter text.[4] It uses recurrent neural network along with LSTM.

1. Ti —>Ci
Where i={1,2…..,n}
T=Tweet
C=Integer representation of Tweet
2. x={x Ti | x is longest sample}
3. E —> Encoder D—> Decoder
4. Hi=[n][k]
Where H=OneHotEncoder
n=No.of character in longest tweet length k=No.of characters in our dictionary
5. E1=Encoder Input Sentence D2=Decoder Input Sample T3=Target
6. Clean the data and append “start” and “end” to T3
7. Convert word —> indexed numbers [using dictionary]
8. Convert word —> fixed length vector using [embedding layer]

After reading the excel sheet, which is updated in the previous step, preprocessing is performed on both text part and the expected summary part which is nothing but output provided by the extractive summary generation process. The preprocessing involves lowering the textcase,removing emoticons,punctuation etc.After preprocessing,a tokenizer is used to transform sequence of words to sequence of integer. [11] Seq2Seq model comprising of Encoder Decoder is used for abstractive summary generation as the input and output sequence to the model vary in length.Encoder is a 3 stacked LSTM.An Encoder Long Short Term Memory model (LSTM) reads the entire input sequence considering one word at a time and then then processes the information at every timestep in order to captures the contextual information present in the input sequence. Decoder is run which gives the output in terms of probability for the next word and the word with highest probability is then chosen.After running the inference phase viz.testing phase for the model, seq2text() function will convert integer sequence to word sequence for summary generation.

We have limited the text(tweet) to 5000 characters and summary to 1000 characters.

A. Working of LSTM

Two main components of LSTM are the cell state nad the hidden state.The hidden state ats as a temporary memory of the neural network.It stores the information about previous data that the network has seen and processed before. The cell state carries the information from the starting to the later stages that which in turn reduces the effect of short term memory.

LSTM (Long-Short term memory) makes use of two main activation function which are as follows:

1. Tanh activation function : It compresses the values so that they are between -1 and 1.This helps in regulating the network.
2. Sigmoid activation function : It compresses the values between 0 and 1.It helps to identify relevant values. The values close to 0 are eliminated and the ones close to 1 are kept for further use.

In the diagram,
C_{i-1} : Previous cell state C_i : New cell State
h_{i-1} : Previous hidden state X_i : Current input

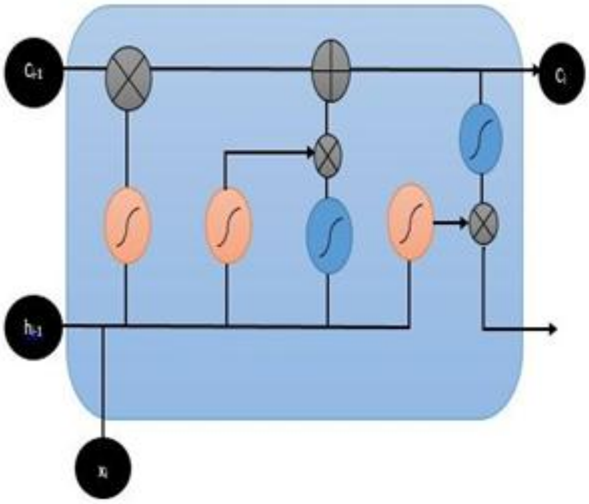


Figure 2. Working of LSTM

Forget Gate : It uses a sigmoid function which compresses the values so that they are between 0 and 1. If the values are close to 0, it forgets the value and if the values are close to 1 then it remembers the values.

The output of forget is a forget vector.
Input Gate : The input gate is used to update the cell state. The previous hidden state (h_{i-1}) and the current input (x_i) pass through the sigmoid function to keep important information and forget the irrelevant information by normalizing values between 0 and 1. The network is
is passed through tanh function. output of tanh function is multiplied with the output of sigmoid function in order to get the information that the next hidden state should hold.

V. RESULT AND DISCUSSION

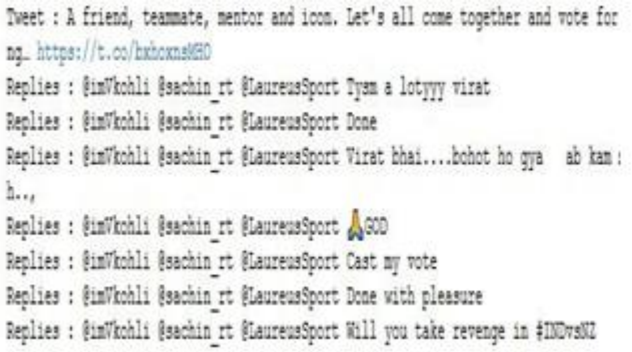
We have checked the precision of summary generated by system by conforming it from general public and considering their feedback while calculating the accuracy.The summary was presented to the users which was generated by our system for a particular user handle, We provided two ways of feedback,one being a poll with “yes/no” option and the other with star ratings.If a ‘no’ is received from the user,it is classified as a “negative feedback”.However,if the user marks ‘yes’,it is furthermore validated with star rating which ranges from 1-5. This further parameter ensures proper validation of the result generated.
In star ratings,2 or less than “2 stars” indicate “unsatisfied opinion” of the user regarding summary even if he has marked ‘yes’ initially.”3 stars” indicate “neutral opinion” and “greater than 3 stars” would indicate a “satisfied opinion.”
The accuracy of the system is calculated on the basis of received feedback with the help of conditional probability.

$$P(Satisfied \mid Yes) = \frac{(Probability\ of\ No\ of\ users\ having\ opinion\ "Yes"\ \&\ "Stars > 3")}{(Probability\ of\ No\ of\ users\ having\ opinion\ "Yes")} * 100$$

For verifying the above,we took feedback of 50 people on the summary generated for a particular userhandle of twitter.Out of which,42 people voted yes & further out of them 35 gave satisfactory ratings(3 stars and above) and the remaining 7 people gave ratings less than or equal to 3.

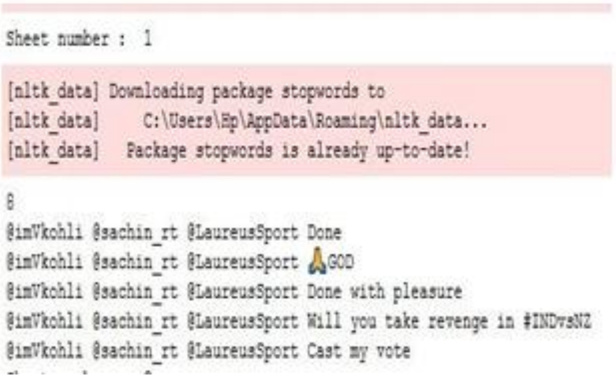
$$P\,(Satisfied\ |Yes) \;=\; \frac{35/50}{42/50} * 100 \quad = 83.33\%$$

regulated by passing the previous hidden state and current input through tanh activation function that normalizes the values between -1 and 1. The output of tanh activationfunction is pointwise multiplied with the output of sigmoid function to decide what important information from the tanh output has to be kept.
Cell state : To formulate the next cell state, the current cell state(C_{i-1}) gets pointwise multiplied with the output of forget gate which is a forget vector that drops values close to 0. The output of the input gate is pointwise added with it to update the cell state to new values that the neural network is expected to remember and which it concludes to be important
Output gate ; It is used to formulate the next hidden state.The previous hidden state(h_{i-1}) and the current input(x_i) is passed through a sigmoid function.The new cell state(C_i)
This survey resulted in an accuracy of 83.33% .
The expected output for a given tweet (retrieved input) along with its top replies in Extractive summarization is as follows:



```
Tweet : A friend, teammate, mentor and icon. Let's all come together and vote for
ng… https://t.co/bxboxnsMHO
Replies : @imVkohli @sachin_rt @LaureusSport Tysm a lotyyy virat
Replies : @imVkohli @sachin_rt @LaureusSport Done
Replies : @imVkohli @sachin_rt @LaureusSport Virat bhai....bohot ho gya   ab kam :
h..,
Replies : @imVkohli @sachin_rt @LaureusSport 🙏GOD
Replies : @imVkohli @sachin_rt @LaureusSport Cast my vote
Replies : @imVkohli @sachin_rt @LaureusSport Done with pleasure
Replies : @imVkohli @sachin_rt @LaureusSport Will you take revenge in #INDvsNZ
```

Figure 3. Extracted Tweets And Replies



```
Sheet number :  1

[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Hp\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!

8
@imVkohli @sachin_rt @LaureusSport Done
@imVkohli @sachin_rt @LaureusSport 🙏GOD
@imVkohli @sachin_rt @LaureusSport Done with pleasure
@imVkohli @sachin_rt @LaureusSport Will you take revenge in #INDvsNZ
@imVkohli @sachin_rt @LaureusSport Cast my vote
```

Figure 4. Extractive Output

VI. CONCLUSION

Summarization systems that are currently existing in the market make use of either statistical approaches or linguistic approaches. Statistical techniques begin with basic features such as term frequency (TF-IDF) and gradually extend to positional features and contextual features in order to ensure high quality summary. The linguistic techniques rely on semantic analysis and adopt Lexical databases to find the association between textual units. This technique generates cohesive summary as compared to statistical techniques using low level features. To achieve benefits of both these approaches, our system makes use of a hybrid approach including statistical as well linguistic techniques. Our system generates both extractive as well abstractive summary.Using this approach, our system generates a summary of the respective user’s activity monthly along with the summary of the replies given to the user in the past 7 days.

FUTURE SCOPE

- **Relegate linguistic restriction:** Currently system can generate summary for only “English” text. System can be extended to generate summary for other language text as well if online lexical database for other languages are available.
- **Product Review summary:** Proposed System can generate summary for comments on twitter posts. It can be extended to generate summary for product reviews on various e-commerce sites.
- **Extension to other social media application:** It is also possible to generate summary for comments on posts of other social media applications like instagram, hike, facebook etc.

REFERENCES

1. Koustav Rudra,Siddhartha Banerjee,"Summarizing Situational Tweets in Crisis Scenario", (2016)
2. XiaoHua, LI YiTong, WEI FuRu, ZHOU Ming LIU"Graph-based Multi-tweet Summarization Using Social Signals" (2016)
3. Hirao Tsutomu, Nishino Masaaki, Yoshida Yasuhisa, Suzuki Jun,Yasuda Norihito, and Nagata Masaaki, "Summarizing a Document by Trimming the Discourse Tree", IEEE/ACM Transactions On Audio, Speech, And Language Processing(2015)
4. Sarda A.T. and Kulkarni A.R., "Text Summarization using Neural Networks and Rhetorical Structure Theory", International Journal of Advanced Research in Computer and Communication Engineering(2017)
5. Renjith S.R, Sony P, "An Automatic Text Summarization for Malayalam Using Sentence Extraction", Proceeding of 27th IRF International Conference(2015
6. Subramaniam Manjula, Dalal Vipul, " Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method." IRJET(2015)
7. N.Moratanch and S.Chitrakala, "A Survey on Extractive Text Summarization", IEEE International Conference on Computer, Communication, and Signal Processing(2017)
8. Arpita Sahoo and Dr.Ajit Kumar Nayak, "Review Paper on Extractive Text Summarization(2018)
9. M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut. 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. ArXiv e-prints (2017)
10. Koustav Rudra, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, Prasenjit Mitra"Summarizing Situational Tweets in Crisis Scenario"(2016)
11. Ankit Kumar, Zixin Luo, Ming Xu,"Text Summarization using Natural Language Processing"(2018)
12. Alexander M Rush, Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization", 2015.
13. Z. J. Fu, X. L. Wu, Q. Wang, and K. Ren, "Enabling central keyword-based semantic extension search over encrypted outsourced data, "IEEE transactions on information forensics and Security, vol. 12, no. 12, pp.2986-2997, 2017.
14. C. Chen, X. J. Zhu, P. S. Shen et al., "An efficient privacy-preserving ranked keyword search method," IEEE Transactions on Parallel and Distributed Systems, 2015.
15. Broenlee, J. "A Gentle Introduction to TextSummarization".March 02, 2018.
16. Dalal V.& Malik, L. G. ,"A survey of extractive and abstractive text Summarization techniques",In Emerging Trends in Engineering and Technology (ICETET) ,2018.
17. Nallapati, R. Zhou, B.,Gulcehre., & Xiang, B. "Abstractive Text Summarization Using sequence-to-sequence RNN's and Beyond" (2016)
18. Radhakrishnan, P. "Attention Mechanism Network" Hacker Noon(2017)

AUTHORS PROFILE

Ayushi Gupta, student of BE, Computer Engineering,MITCOE
"Fake Email & Spam Detection Using Naïve Bayes & User Feedback Approach"-Springer,ICCSA 2019
"Fake News classification on Twitter using Flume, N-gram analysis and Decision Tree machine learning technique"-Springer,ICCSA 2019
"Human Gait Analysis based on Decision Tree, Random Forest and kNN Algorithms" -ICCET, Jan 2020.Her area of interests are Data Science ,analysis & its applications

Devyani Keskar, student of BE,
Computer Engineering, MITCOE, "Fake News classification on Twitter using Flume, N-gram analysis and Decision Tree machine learning technique"-Springer, ICCSA 2019
"Fake Email & Spam Detection Using Naïve Bayes & User Feedback Approach"-Springer,ICCSA 2019
Her area of interests are Big Data & programming

Madhur Firodiya , student of BE, Computer Engineering, MITCOE, His area of interests are Database and its management

Siddhi Hagawane, student of BE, Computer Engineering, MITCOE, Her area of interests are in Cyber Security and its application