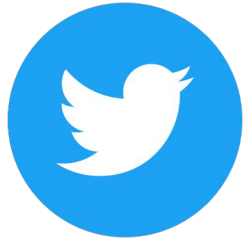# Sentiment and Emotion Analysis of Indian Union Budget 2022-23 on Social Media

Negative Neutral Positive

GROUP 24

ADITYA DIXIT
MT21004

MOHNISH BASARKAR
MT21052

SHIVANG KANWAR
MT21084

SHUBHAM AGGARWAL
MT21090

1

PROBLEM STATEMENT

Finding the sentiments and emotions of Indian citizens regarding the current financial year budget i.e. 2022 - 23.

2

LITERATURE REVIEW

- All papers have created their dataset by scraping tweets from twitter.

- Various hashtags like #unionbudget, #budget2020 are used for scraping data.

- Then preprocessing of tweets are done to remove any garbage data and make tweets ready for further operations.

- In sentibase paper, authors have used linear SVM to classify tweets into positive, negative and neutral.

- Sentiment analysis on the union budget paper uses different process and first makes Term Document matrix to find frequency of each word and then uses this matrix to find relevant words for doing sentiment analysis.

- In most of the papers, sentiments are classified using TextBlob library of python which classifies tweets on the basis of subjectivity and polarity.

- In paper Sentiment and emotion analysis of indian budget various techniques are used like lexical analysis, statistical approach for sentiment score. They give their result by showing 8 emotions and percentage of people having these emotions.
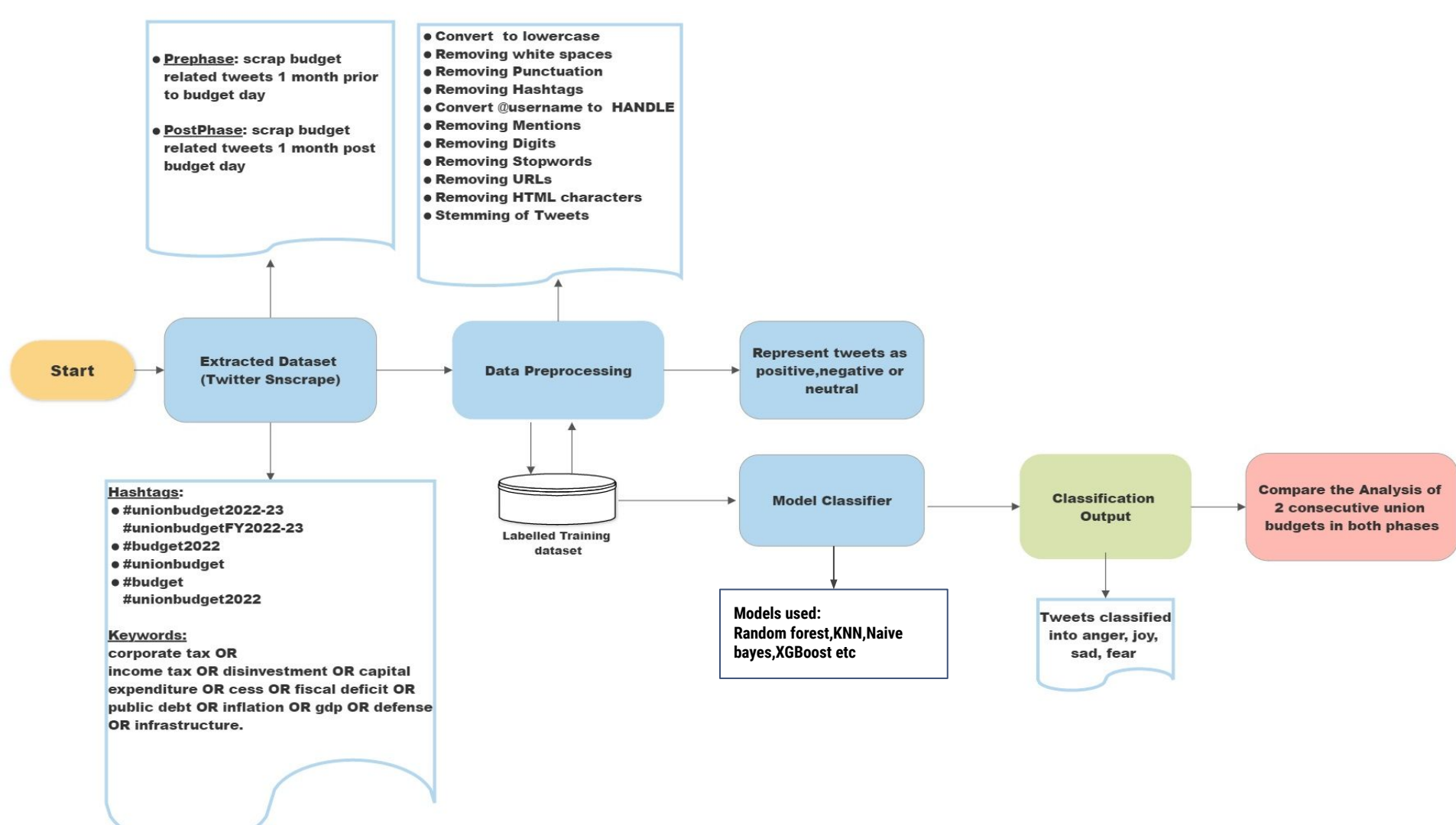
3

CONTRIBUTION TO THE EXISTING WORK

- Along with english tweets we have also accumulated hindi tweets and comments from twitter and other platforms to give **bilingual support**.

- Apart from twitter we have also collected data in form of comments from youtube where union budget was streamed live , this not only helped us to have more data for analysis but also gave information about sentiments of people on platforms like youtube i.e. **multi-platform**.

- Unlike other research works which mainly focussed on either sentiment analysis or emotion analysis ,our research lay stress on both aspects at the same time.

- We have divided our research into 2 phases that are pre-budget phase and post-phase so we can compare change in emotions and sentiments of people for the two phases.
- We have also compared analysis of this year with the analysis of previous year so that change in emotions and perception of the people can be looked into.
- For machine learning approach we didn't just stick to one machine learning model but rather we went on to compare ten different models with our base model ie Random Forest Classifier to come up with the best one which was XGBoost.

4

ARCHITECTURE

Prephase: scrap budget related tweets 1 month prior to budget day

PostPhase: scrap budget related tweets 1 month post budget day

- Convert to lowercase
- Removing white spaces
- Removing Punctuation
- Removing Hashtags
- Convert @username to HANDLE
- Removing Mentions
- Removing Digits
- Removing Stopwords
- Removing URLs
- Removing HTML characters
- Stemming of Tweets

Start

Extracted Dataset (Twitter Snscrape)

Data Preprocessing

Represent tweets as positive, negative or neutral

Labelled Training dataset

Model Classifier

Classification Output

Compare the Analysis of 2 consecutive union budgets in both phases

Hashtags:
- #unionbudget2022-23 #unionbudgetFY2022-23
- #budget2022
- #unionbudget
- #budget #unionbudget2022

Keywords:
corporate tax OR income tax OR disinvestment OR capital expenditure OR cess OR fiscal deficit OR public debt OR inflation OR gdp OR defense OR infrastructure.

Models used:
Random forest, KNN, Naive bayes, XGBoost etc

Tweets classified into anger, joy, sad, fear

TOOLS, TECHNOLOGY & LIBRARY USED

- SPYDER IDE
- JUPYTER NOTEBOOK
- NATURAL LANGUAGE PROCESSING (NLP)
- SNSCRAPE LIBRARY (for Scraping Tweets from Twitter)
- PYTHON 3.x
- TextBlob library (Finding Polarity of tweets)
- "Googletrans" library (Converting Hindi tweets to english)
- "MATPLOTLIB" library (Building graphs)

Different Classifiers Used

- RANDOM FOREST
- MULTINOMIAL NAIVE BAYES (MNB)
- SUPPORT VECTOR CLASSIFIER (SVC)
- LINEAR SUPPORT VECTOR CLASSIFIER (LinearSVC)
- K NEAREST NEIGHBOURS (KNN)
- XGBOOST (XGB)
- LOGISTIC REGRESSION (LR)
- STOCHASTIC GRADIENT DESCENT (SGD)
- ADABOOST
- DECISION TREE CLASSIFIER (DTC)
- GRADIENT BOOSTING CLASSIFIER (GBC)

5

CODE FLOW

- Preprocessing of data which is scrapped from twitter using standard preprocessing techniques like removing punctuations , stop words , replacing hashtags ,removing urls , replacing repeating words and stemming.
- Similarly scrapped and preprocessed hindi tweets and comments from other social media platforms in order to have bilingual support with the help of "*googletrans*" library as well as more spatial data for further analysis.
- After preprocessing step generated sentiments of each tweets using text blob library .
- Using Tf-idf vectorizer we converted our raw data into tf-idf features so that we can train models using that data.
- Now by splitting above raw data into training and validation we trained different models and analysed them on different scores like F1 score , precision, recall and accuracy then also studied sentiment prediction given by them.
- For emotion analyses of tweets we used a standard labelled data to train our different models and then using these models we predicted sentiments of tweets that we scrapped initially.

6

ANNOTATING DATA

- For annotating tweets/responses into positive, negative and neutral sentiments. We have use **TextBlob** library of python.


Comparison between different phases considered for both financial years
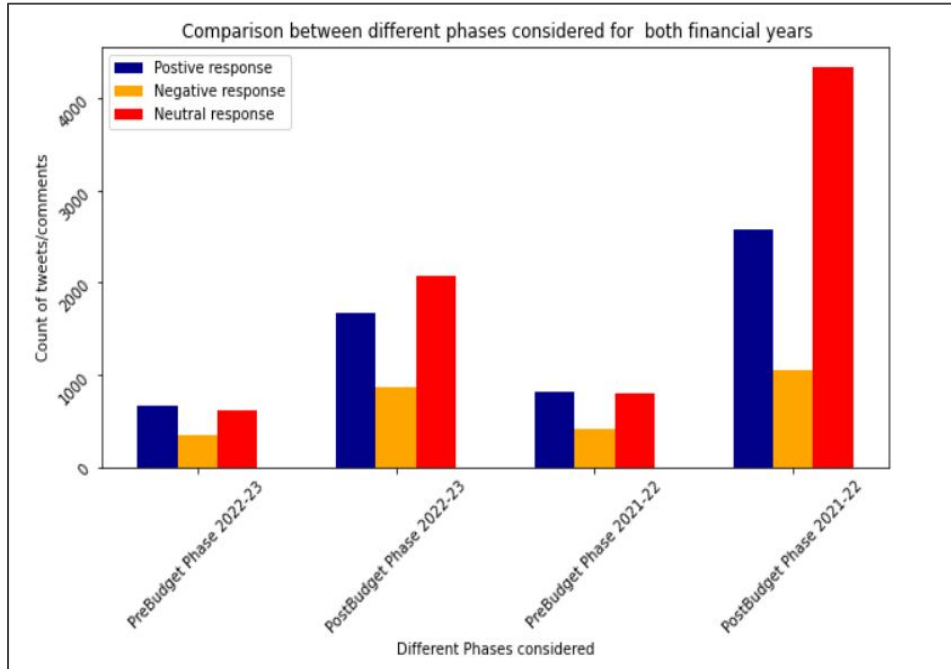
Phases	Positive	Negative	Neutral
PrePhase 2022-23	671	340	620
PostPhase 2022-23	1699	857	2066
PrePhase 2021-22	808	412	805
PostPhase 2021-22	2571	1051	4345

- For annotating tweets/responses into various emotions like joy, anger, disgust, shame, sadness etc. We have use **Random Forest Classifier** in python.
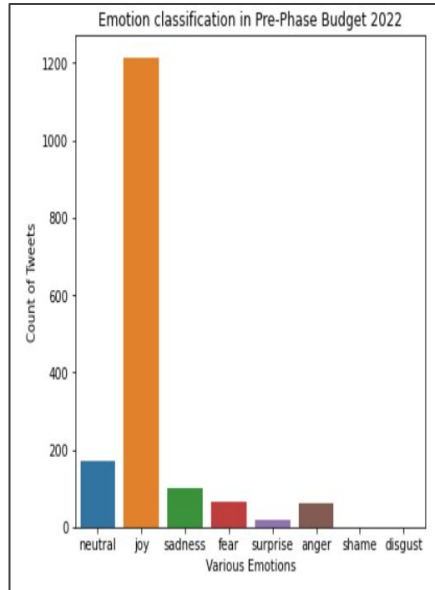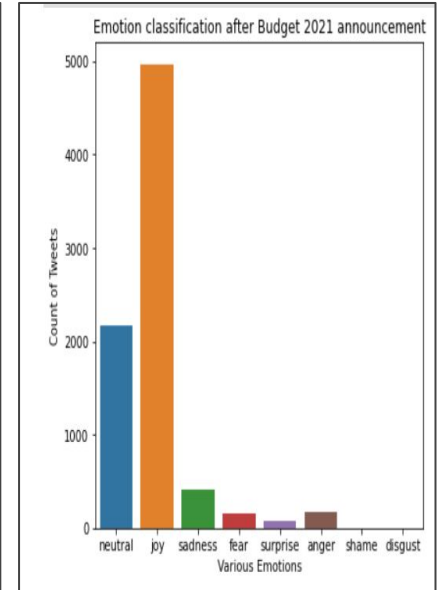

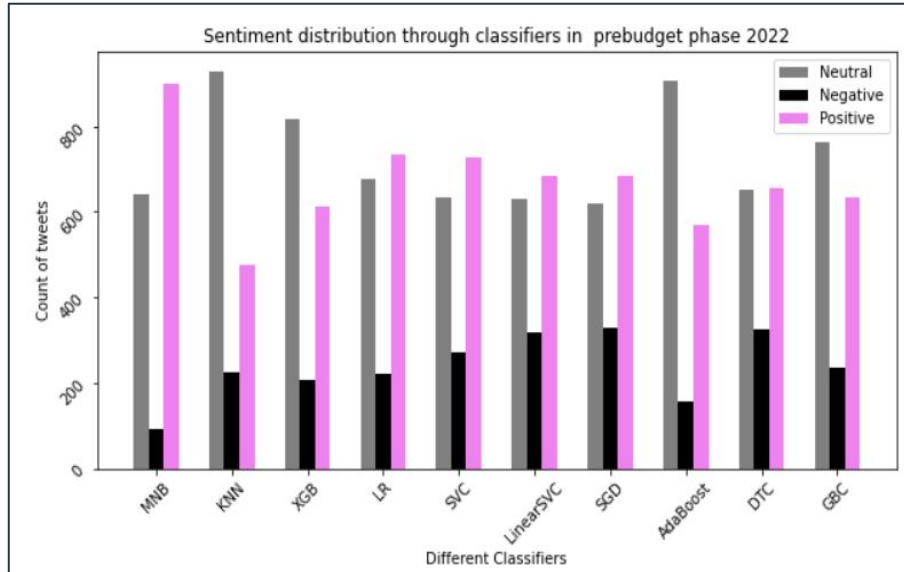Pre-Phase 2022-23
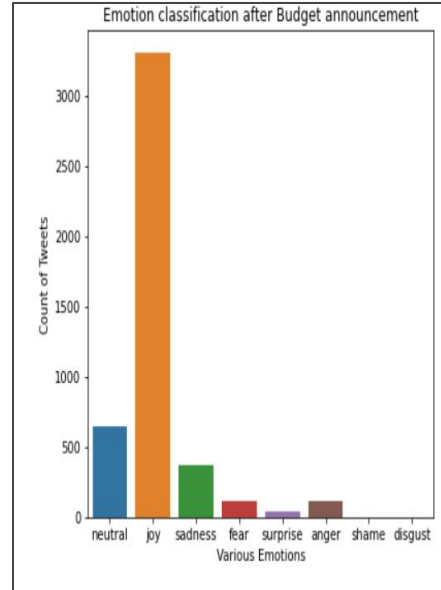
Post-Phase 2022-23
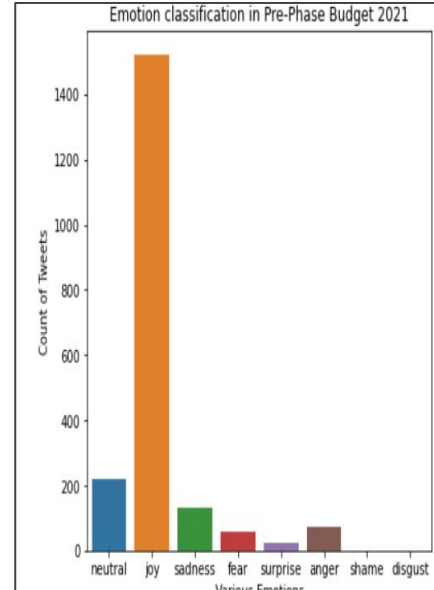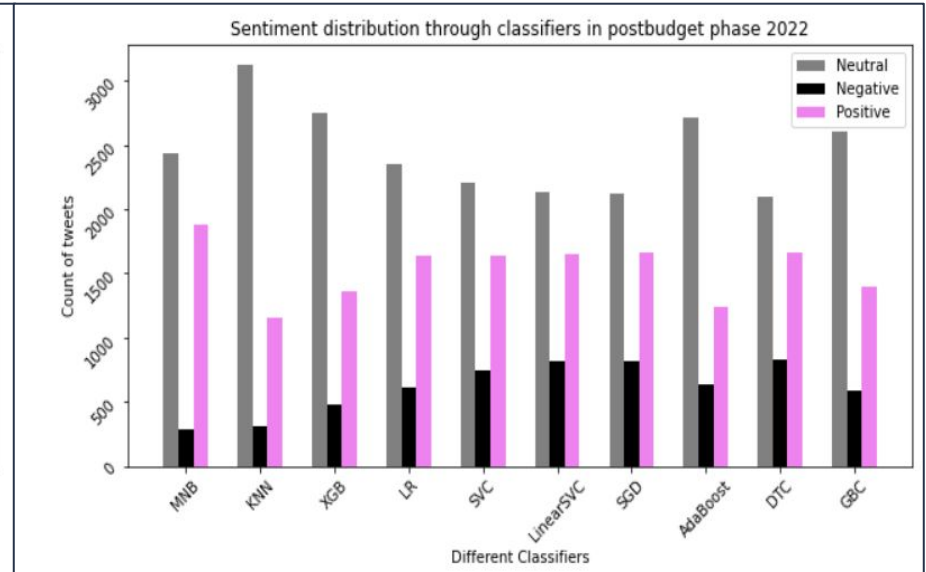
Pre-Phase 2021-22

Post-Phase 2021-22

7

COMPARISON USING DIFFERENT CLASSIFIERS
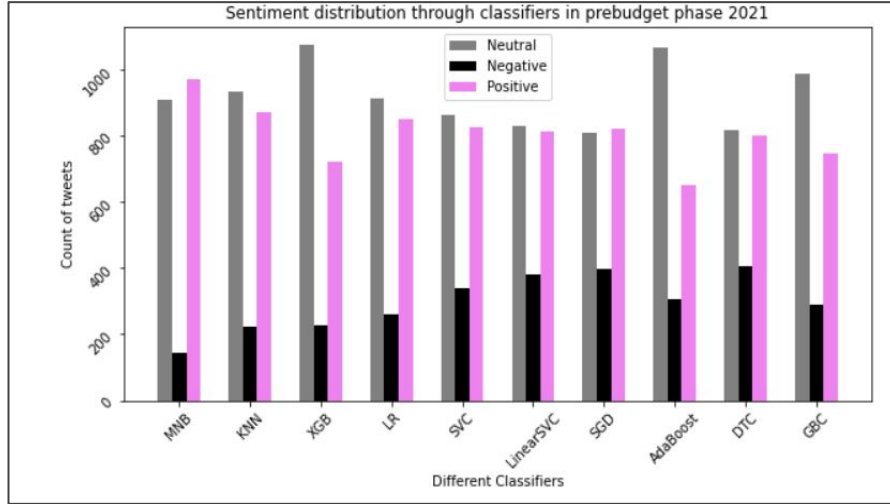
● Classification of tweets into positive, negative and neutral by different classifiers.
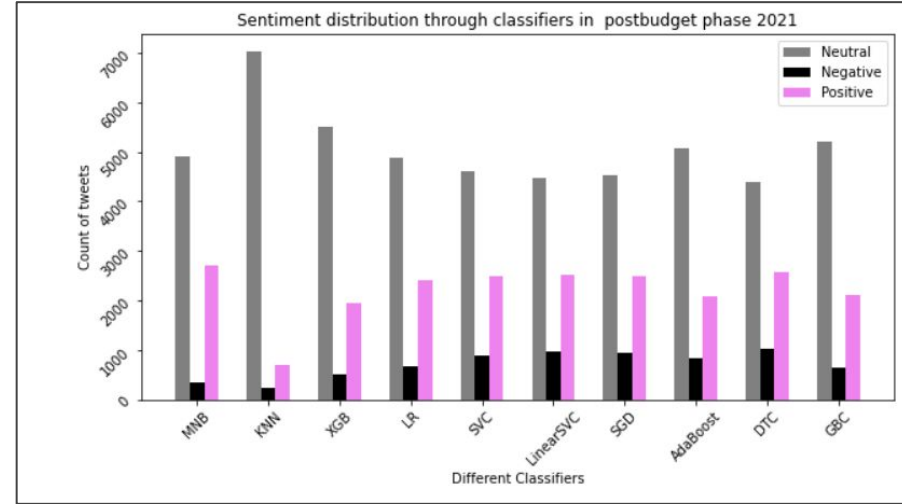

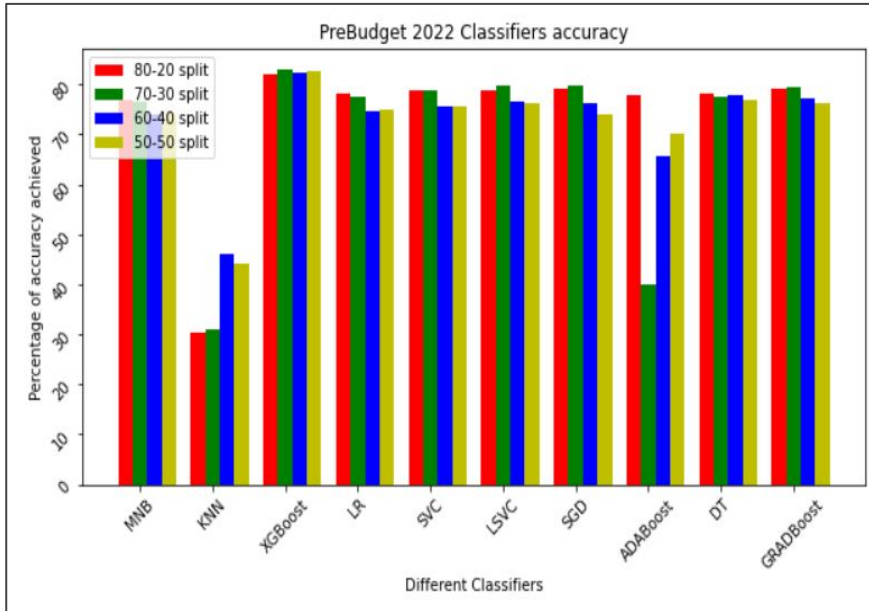Pre-Phase 2022-23

Post-Phase 2022-23

Sentiment distribution through classifiers in prebudget phase 2021

Pre-Phase 2021-22

Sentiment distribution through classifiers in postbudget phase 2021
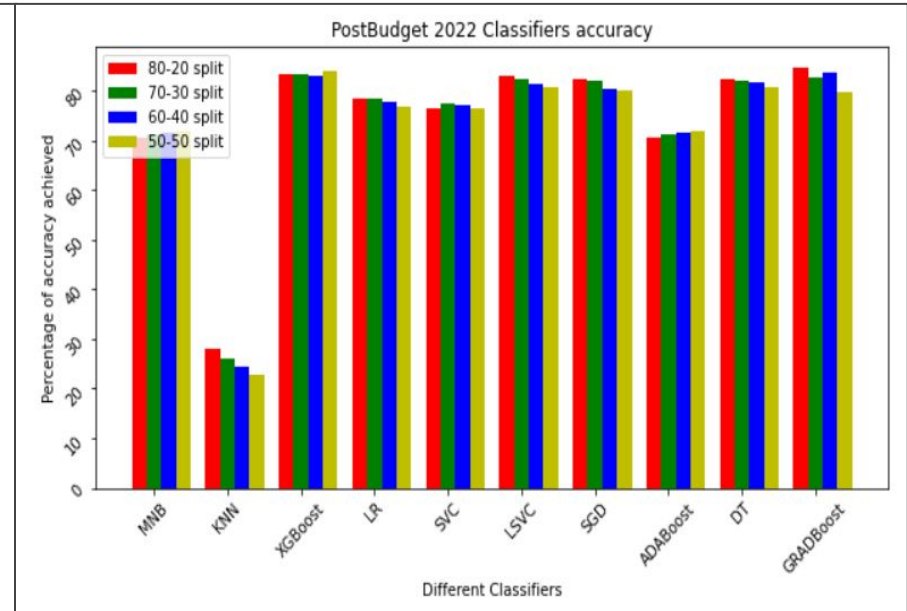
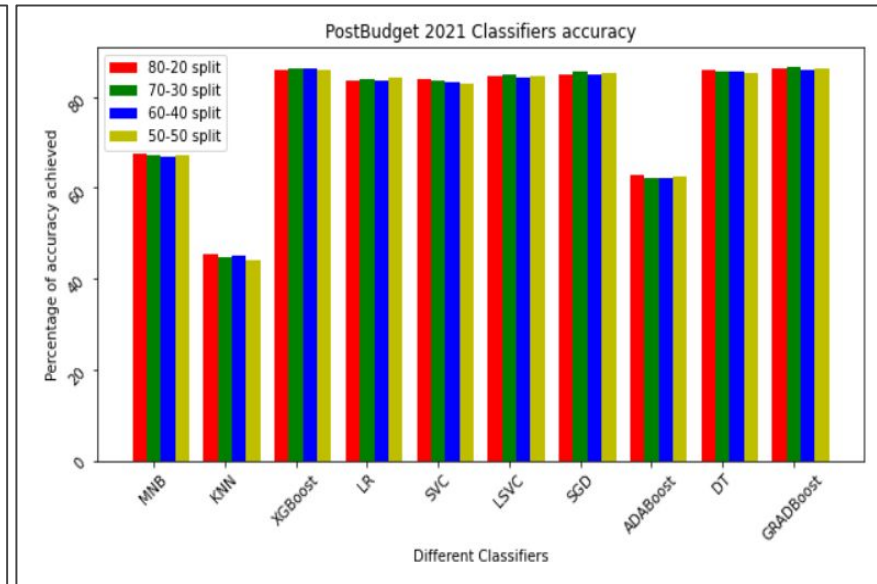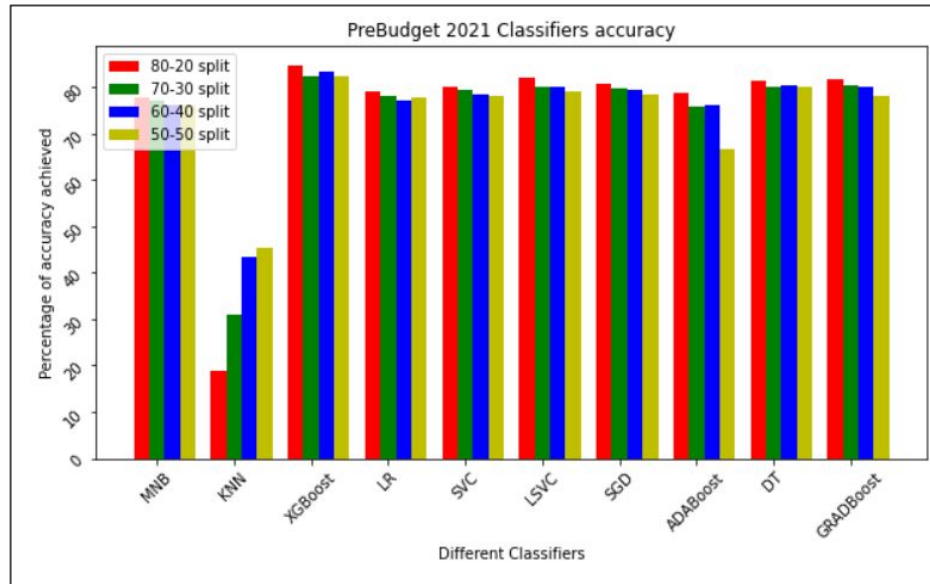Post-Phase 2021-22

- Classification of emotions by different classifiers on the basis of different splits.

Pre-Phase 2022-23   Post-Phase 2022-23

Pre-Phase 2021-22

Post-Phase 2021-22

XGBoost comes out to be the best classifier with an accuracy of 83.10% for 2022-23 and 84.57% for 2021-22.

8

RESULTS

➢ **Sentiment analysis** results were as follows:

Sentiments	Pre-Budget 2022-23	Post-Budget 2022-23	Pre-Budget 2021-22	Post-Budget 2021-22
Positive	671	1669	808	3571
Negative	340	857	412	1051
Neutral	620	2066	805	4335

- In both the phases for both financial years, the number of positive tweets/comments were relatively higher than that of negative tweets/comments indicating positive attitude towards the budget.

- For the Budget year 2022-23, percentage of negative tweets remained same for pre and post budget. However, percentage of positive tweets dropped from 41.1% during pre-budget phase to 35.4% after the declaration of budget indicating some sort of unhappiness.

➢ **Emotion analysis** results were as follows:

● On performing the emotion analysis, maximum number of tweets were classified as 'Joy' with 79% and 69.6% of the total tweets and comments for the financial year 2022-23 and 2021-22 respectively. However, the number of tweets and comments from youtube dropped for the year 2022 as compared to that of year 2021.

● Following classification models were used in emotion analysis: Random Forest(as baseline), MNB, KNNClassifier, XGBClassifier, LogisticRegression, SVC, LinearSVC, SGDClassifier, AdaBoostClassifier, DecisionTreeClassifier, GradientBoostingClassifier.
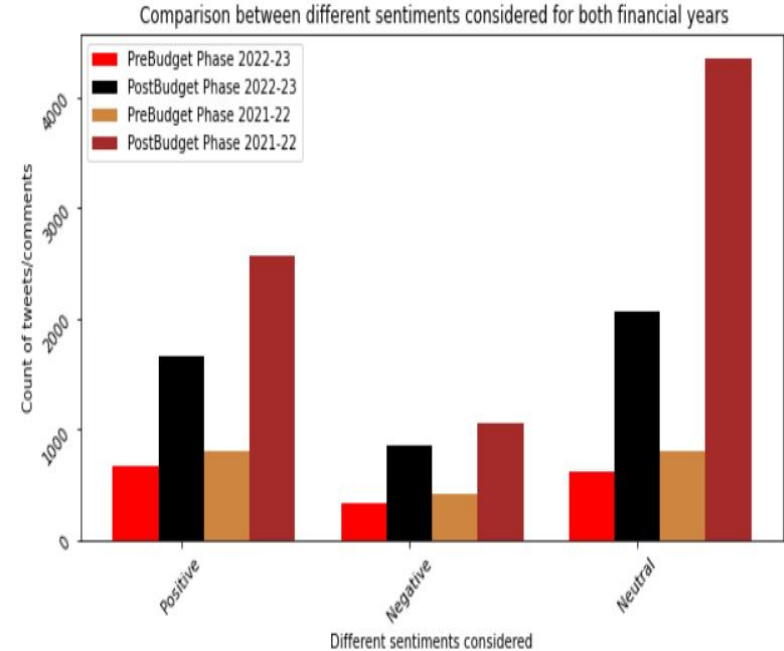
Out of all these, XGBoost gave the highest accuracy of 84.5% and 83.10% for year 2021 and 2022 respectively.
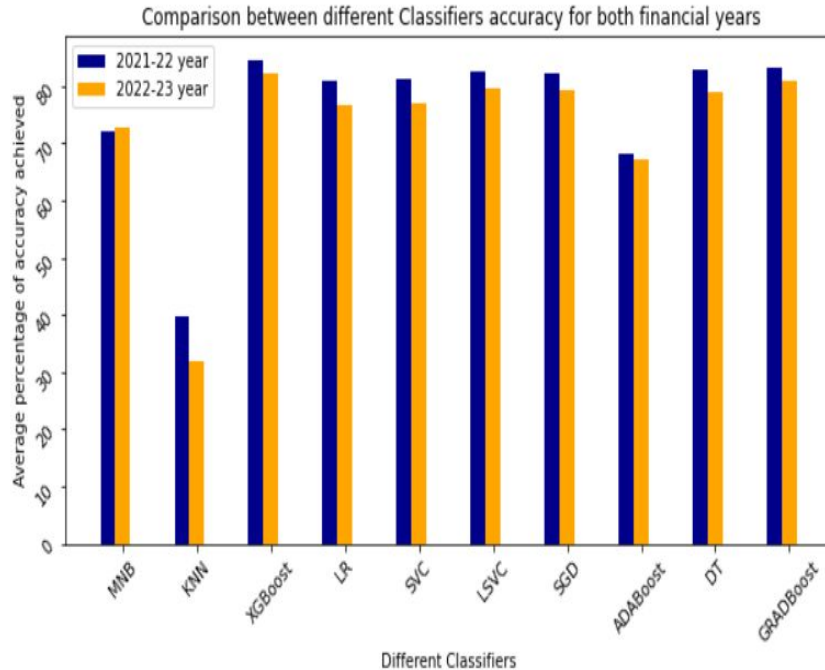
9

ANALYSIS

➢ **In terms of Sentiments**

1. Count of tweets in the post-budget phase 2021-22 is more in number in all 3 types of sentiments which means that policies in that budget are hot topics to discuss.

2. For both financial years, post-budget tweets are more in number which is obvious to discuss after its declaration.

3. But it is interesting to note that pre-budget phase tweets are less but are present. This indicates that people are curious about the coming budget every year to know the benefit they get in an incoming budget.


Comparison between different sentiments considered for both financial years

Comparison between different Classifiers accuracy for both financial years

➢ **In terms of emotions**

1. **XGBoost** classifies the emotions into different categories with an **highest accuracy of 83.10% for 2022-23 and 84.57% for 2021-22**.
2. **KNN** classifies the emotion into different categories with a **lowest accuracy of 40%**.
3. This analysis is an average of all 4 types of splits in both the phases for each financial year.

- Since Count of tweets decreases in every emotion. It indicates that citizens are less inclined towards current financial year budget i.e. 2022-23 as compared to last year budget 2021-22.
- But still joy emotion was dominating in both financial year. Means that people are satisfied with policies of both budget.
- Very less/minimal indication of sadness emotion is observed which may be due to affecting certain policies to specific segment of society.

Comparison between different emotions through XGBoost classifier for both financial years
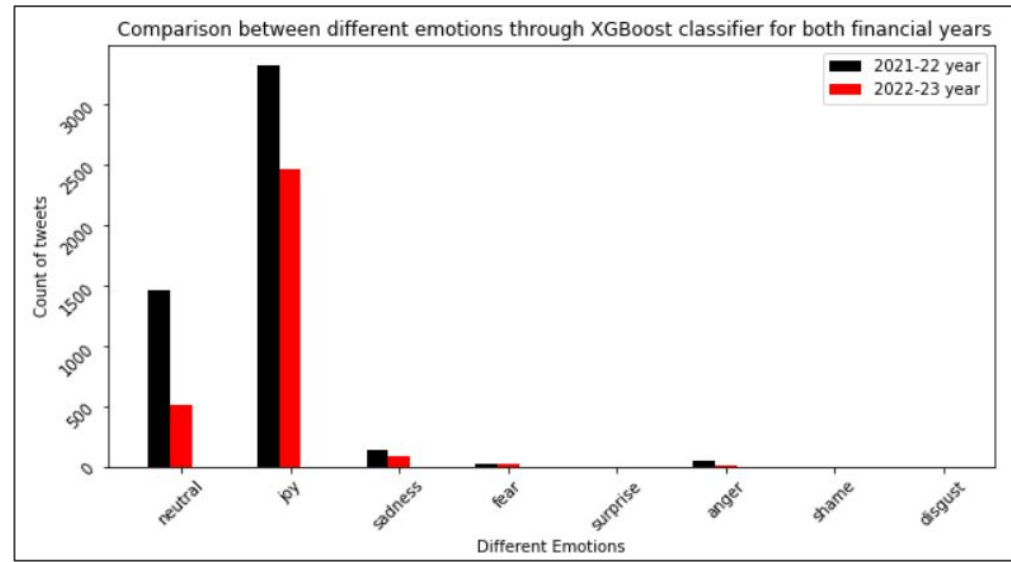
Different Emotions	Count of tweets achieved for 2022	Count of tweets achieved for 2021
neutral	515	1456
joy	2466	3323
sadness	93	145
fear	23	23
surprise	4	5
anger	11	45
shame	0	0
disgust	0	0

10

LIMITATIONS

- **Polarity**: words like 'hate' and 'love' have polarity scores as -1 and 1 respectively. There are sometimes in-bw conjugation of words like 'not so bad' which gets left out,diluting the sentiment score.
- **Sarcasm & Irony**:using irony and sarcasm to express negative sentiments makes it difficult to identify true context,resulting into higher positive sentiments.
- **Multilingual data**:translating into base language like English, the meaning and context might get lost. Solution is to train model in each language manually, but a time-consuming process.
- **Non-textual data**:gifs,images and videos are also used to express sentiments on social media.

11

CHALLENGES

- Collecting the decent amount of data to create a dataset for analysis.
- Preprocess the raw tweets and comments so that it becomes suitable for processing and analysis.
- Giving support to the multiplatform and multilingual functionality.
- Explore about different classification and ML techniques.
- Design the architecture and flow of complete project analysis.

12

FUTURE SCOPE

- Comparing tweets/comments of more than 1 previous year, so as to analyze how the trend of public opinion changes related to Indian Budget over the period of time.
- Classification of tweets/comments into major sectors of budget like service, defence, agriculture etc and perform sentiment and emotional analysis categorically and analyze the impact of budget sector-wise.
- Budget analysis can be done for any country in any language.
- Collect tweets/comments from other platforms like reddit, facebook etc.
- Improve accuracy and efficiency of models used.

13

CONCLUSION & DISCUSSION

- This project will help in decreasing the communication gap between the government and the public.

- With the help of analysing sentiment of a large mass we can show the inclination/interest of crowd towards the budget.

- This will help the government to make such policies more in number for the betterment of the people of the country.

- If such things happens then public will also show interest in giving their opinion freely on twitter which can create an engaging relationship between the public and government of a country.

- Classification of budget into depth encourages citizens to explore it and gain knowledge about various aspects of budget which might get useful for them in future.

14

REFERENCES

1. Dr V Uma Maheswari Abhinand G. 2022. Corroboration of Twitter Sentiment Analysis and Event Analysis of Indian Budget 2022 on Bitcoin Market. (April 2022), 12.

2. Seema S. Kawathekar Bharat Naiknaware. 2018. Peoples Opinion on Indian Budget Using Sentiment Analysis Techniques. (2018), 5.

3. Satarupa Guha, Aditya Joshi, and Vasudeva Varma. 2015. Sentibase: Sentiment Analysis in Twitter on a Budget. (07 2015), 590–594.

4. Sheetal Mahendher, Toshith Sastry, Yashus Gopal, and Rohith M S. 2021. Sentimental Analysis on the Union Budget, India-2020. 10 (02 2021), 14–21.

5. Manpreet Singh Rupinder Kaur, Rajvir Kaur and Dr. Sandeep Ranjan. 2020. Twitter Sentiment Analysis of the Indian Union Budget 2020. 29 (2020), 8.

6. http://www.ijsrcsams.com/images/stories/Past_Issue_Docs/ijsrcsamsv7i4p149.pdf

7. Moonis Shakeel and Vikram Karwal. 2016. Lexicon-based sentiment analysis of Indian Union Budget 2016–17. (july 2016), 299–302.

8. Monali Waghmare and Sachin Deshmukh. 2018. Sentiment and Emotion analysis on Indian Budget. 7 (july 2018), 4.

PROJECT RESOURCE LINKS

- GitHub Link :

  https://github.com/subuu/IR2022_FinalProject_24

- Report Link :
  https://github.com/subuu/IR2022_FinalProject_24/blob/main/IR2022_Project_Report_24.pdf

- Code File :
  https://github.com/subuu/IR2022_FinalProject_24/blob/main/Final_IR_Project_Group_24%20.ipynb

THANKS!!