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

Aditya Dixit
aditya21004@iiitd.ac.in
IIIT Delhi
India

Mohnish Basarkar
mohnish21052@iiitd.ac.in
IIIT Delhi
India

Shivang Kanwar
shivang21084@iiitd.ac.in
IIIT Delhi
India

Shubham Aggarwal
shubham21090@iiitd.ac.in
IIIT Delhi
India

KEYWORDS

Union Budget,Sentiment Analysis,Twitter,YouTube,Emotions,Text Mining

1 PROBLEM STATEMENT

Analyzing the emotions and sentiments of Indian citizens towards Union Budget for financial year 2022-23 using social media platforms like twitter and youtube.

2 MOTIVATION

Social media not only facilitates conversations, creation, and sharing of information, it also offers a huge corpus for different types of analysis like Misinformation Analysis, Sentiment Analysis, and so on. Sentiment Analysis in social media can be useful to understand the opinion of people about a variety of topics, for example, Governments can analyze the public's opinion over covid vaccines and can take necessary actions. Companies can study consumers' sentiment towards their product and develop business strategies, political parties can predict their winning chances in the upcoming elections by monitoring voters' opinions about them, and so on. Sentiment analysis on Twitter and youtube can be used to monitor the public's opinion over the Indian Union Budget 2022. This will help Government get the feedback of the public on the Union Budget whether they liked the central theme of the budget or felt disappointed. It will also allow the government to make changes in the policies that will add to the growth of the country.

3 PROBLEM IMPORTANCE

Social media acts as a platform to for people to express their opinions on various topics. A wide variety of views related to Indian Union Budget, often colored by emotions which changes with nearing budget session and post its declaration are discussed online. Twitter and youtube has now become a major platforms to express an opinion regarding each and every event happening around the world, hence both of them can be used as a major source to draw opinions or sentiments of people regarding any hot topic. As many things in this world majorly depend upon perception as well as sentiments of people, hence sentiments of people towards any work done by the government can be a very strong driving force, thus this problem holds importance in designing policies according to the sentiment of the masses which would help in making efficient policies in future.

4 LITERATURE REVIEW

In 2015 [3], Satarupa Guha, Aditya Joshi, Vasudeva Varma in their paper Sentibase: Sentiment Analysis in Twitter on a budget, data from Twitter is collected and operated upon. In preprocessing, authors have expanded all acronyms to their original form and change number to strings. The next step is vocabulary generation in which they have given unique IDs to all the words occurring in the data. All the hashtags are given single unique IDs as they are not important for sentiment analysis. Next, for the classification step, they have classified tweets into subjective(positive and negative) and objective(neutral). Various features of tweets were analyzed before classifying them. After this, a linear SVM classifier is trained using the python scikit-learn library for classification. The paper's accuracy in classifying tweets is checked by calculating the F1 score, precision, recall for the classifier.

In 2016 [6], Moonis Shakeel and Vikram Karwal have aimed to estimate the sentiment score of India's Union Budget document for 2016-17, whose central theme as anticipated, was rural India. In this, the Budget document was converted to a .txt file format for text data retrieval. R language was used for the analysis. The text file data was preprocessed and converted to a document term matrix displaying the frequency matrix of the text data and the words in it were sorted in decreasing order of their occurrence. Then the association between words that related to the words which were anticipated in the budget 2016-17 was carried out. The word cloud representing the most occurring words was created. The analysis pointed out that the words related to taxation had the highest frequency and deviated from the main focus of the Union Budget as anticipated. Inequality. The sentiment displayed by the budget document can be safely considered neutral. It mostly talks about taxation and less about other aspects of the economy. This paper suggested that the social and economic growth of India is regulated by regulating taxes rather than through other more relevant means.

In 2018 [7], the sentiment and emotional analysis were done by the students of Babasaheb Ambedkar University, Maharashtra. They aim to understand the relevance of public opinion towards the union budget of India 2018 through the large online data resource Twitter. They have used ML techniques, lexicon analysis such as Bing Liu and NRC, and statistical approaches for sentiment score calculation and detection of the sentiment using Twitter tweets. Studied strategies like supervised, unsupervised, lexicon, similarity metric, SVM, etc, which were used by other researchers.

They have completed their study using R programming tools and packages. Around 10k tweets were collected using hashtags like ‘Budget2018’, ‘UnionBudget2018’, ‘BudgetSession2018’. They have used 8 emotions to express their analysis i.e. joy, trust, anger, disgust, anticipation, surprise, sadness, and fear. After performing sentence-level sentiment analysis they concluded that due to the positive polarity shown in text processing, the public inclination is towards government policy.

In 2018 [2], researchers accumulated 3 consecutive year tweets related to the Indian Budget for 2016, 2017 and 2018. Twitter is used to make a large corpus of dataset. This study aims to analyze the opinion of the Indian crowd and tries to infer some technical sense from that budget data. Out of document, sentence and feature level sentiment analysis, sentence level classifier is used to determine the polarity of sentiment i.e. positive, negative and neutral. To implement this sentiment analysis they have used R programming open source tools. Statistical, grammar and machine learning based methodologies used. They have used “Score = No. of positive words - No. of negative words”. If score > 0 then it is positive sentiment , else if score < 0 then it is negative sentiment otherwise it is neutral sentiment. They have concluded that best polarity is observed in the tweets of budget 2018.

In 2020 [5] , Twitter sentiment analysis was done on Indian union budget 2020 by students of GNA University, Phagwara aims to study a public reaction on 6000 real-time tweets using the hashtag “Budget2020”. For this, they have used tweepy which is a Twitter API to consolidate their study results. All the preprocessing stuff has been performed to attain higher accuracy. They finalize their sentimental score based on the concept of polarity and subjectivity with an overall positive score of +149.3387. To assign subjectivity and polarity to the most frequent words in the dataset they have used the Textblob library of python. Finally, they concluded that public reaction is in favor of government policy due to the high positive score achieved from the sentimental analysis.

In 2021, in the paper [4], Twitter was chosen to extract data from, all tweets were downloaded which were posted using the #unionbudget hashtag over a period of one month coinciding with the announcement of the budget. In this paper, data was captured using python and transferred to an excel file and then it was cleaned by removing punctuations, stop words, and neutral words. Words lesser than three letters, numbers, and special characters were removed. In this paper TermDocumentMatrix named function was used to make a document matrix table to depict the frequency of each word, the ten most frequent words are also mentioned in the results. Here sentiment analysis was done using EmoLex (list of words and their association with 8 basic emotions). And two sentiments ie negative and positive, here a dataframe was created where each row of dataframe had a sentence whose sentiment analysis needs to be done and there are ten columns that represent eight emotions and two sentiments. In this, a word cloud was generated using the above methodologies which depicted different words with different sizes depict their relevancy, it was observed that the word “economy” had the highest frequency with maximum use in 427 tweets.

In 2022 in the paper [1], the authors tried to corroborate the results of sentiment analysis with event analysis for the Indian Budget 2022 on cryptocurrencies. They tried to collect the tweets based on keywords like ‘Indian Budget 2022’ and ‘Bitcoin’. They have used supervised Machine Learning algorithms. Tweets were collected 120 days prior to the release of the budget and 10 days after its announcement as the estimation window. They have used “Abnormal normal returns were measured for the constant mean model”. For feature extraction, they have used the TF-IDF strategy and for finding the polarity of tweets they have used the TextBlob library of python. Models like Support Vector Machine, Bernoulli NaÃŕve Bayes, Logistic Regression for comparing accuracies. Logistic regression recorded with the best figures around 81.4 percent. They find that “negative sentiments which were expressed in tweets on BTC-INR were slightly higher than the positive tweets and statistics revealed that the positive sentiments were negated by the negative sentiments.”

5 PLAN OF WORK

As explained in the fig., 1. Dataset was created from Twitter and Youtube video’s comment and replies. Tweets related to budget 2022-23 were scrapped from Twitter using Snscrape library and comments from Youtube videos that live streamed Budget session were scraped. Hashtags related to budget(from Google trends) like ‘Budget2022’, ’BudgetSession2022’ along with keywords like budget, tax, finance, etc were used to extract tweets. Then we performed pre-processing on the data extracted. We performed Sentiment Analysis to represent tweets as ’Positive’ or ’Negative’. Classification of pre-processed data was done to classify into 8 emotional categories using Random-Forest classifier.We will also be using other classification models to label and find accuracies and perform further analysis. Final outcome of the project will be whether the public supports the coming budget policy in a positive way or not. Google collab would be used for the coding part.

6 PROPOSED METHOD

- **Design and Dataset:** We performed our analysis by dividing the timeline into two phases; the Pre-phase starting from 1st Jan 2022 to 31st Jan 2022, and Post-phase starting from 1st Feb 2022 to 28th Feb 2022. We scraped more than 1500 tweets in Pre-phase and more than 4800 in Post-phase using the Snscrape library. For Post-phase, we also scraped approximately 1200 comments and replies from the top 10 Youtube videos that live-streamed UnionBudget22 on 1st Feb 2022. For extracting tweets, the query was created using an ‘OR’ combination of hashtags and keywords related to the Indian Union Budget.

  Query used for scraping tweets : (#unionbudget2022-23 OR #budget2022 OR #unionbudget OR #budget OR #unionbudget2022 OR #unionbudget22-23 OR “corporate tax” OR “income tax”).

  Pre-processing of tweets was carried by converting into lowercase text,removing whitespaces,punctuations,removing all URLs and username,removing # from words,emoticons.Later preprocessed tweets were followed by stemming.
- **Classify pre-labeled data into emotions:** We used a pre-labeled dataset to classify the tweets into one of the 8 emotion categories: ‘neutral’, ‘joy’, ‘sadness’, ‘fear’, ‘surprise’, ‘anger’, ‘shame’ and ‘disgust’.

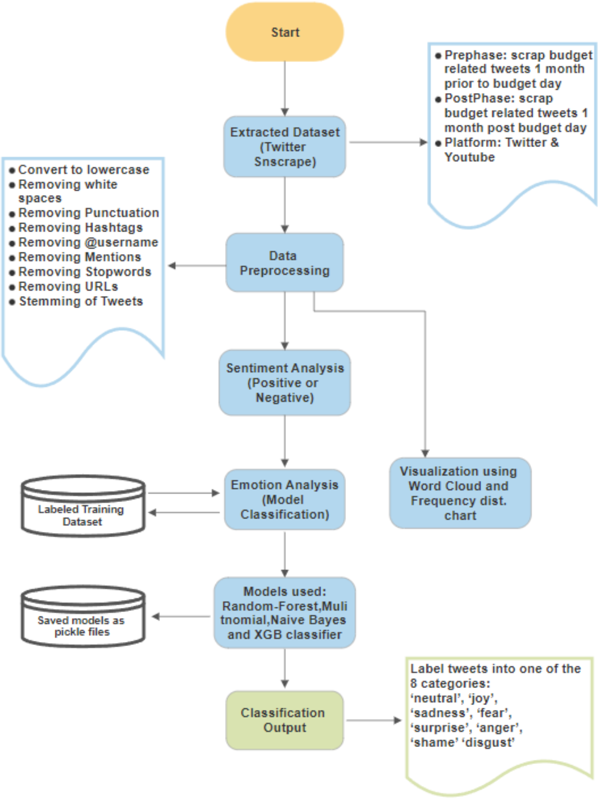
Start
Extracted Dataset (Twitter Snscrape)
● Prephase: scrap budget related tweets 1 month prior to budget day
● PostPhase: scrap budget related tweets 1 month post budget day
● Platform: Twitter & Youtube
Data Preprocessing
● Convert to lowercase
● Removing white spaces
● Removing Punctuation
● Removing Hashtags
● Removing @username
● Removing Mentions
● Removing Stopwords
● Removing URLs
● Stemming of Tweets
Sentiment Analysis (Positive or Negative)
Visualization using Word Cloud and Frequency dist. chart
Emotion Analysis (Model Classification)
Labeled Training Dataset
Models used: Random-Forest,Mulitnomial,Naive Bayes and XGB classifier
Saved models as pickle files
Classification Output
Label tweets into one of the 8 categories: 'neutral', 'joy', 'sadness', 'fear', 'surprise', 'anger', 'shame' 'disgust'

Figure 1: Workflow

- **Pre-budget phase analysis:**

(1) **Sentiment Analysis:** We found the sentiments of each tweet as 'Positive' or 'Negative' using the polarity score of each tweet calculated using the 'SentimentIntensityAnalyzer' tool.

(2) **Emotional Analysis:** We used the pre-labeled dataset to train our Random-Forest classification model. Later, we used this model to classify the tweets and comments scraped from twitter and youtube into one of the 8 emotional categories.

(3) **Visualization using Word cloud and Frequency distribution:** We created the word cloud to visualize the most frequently occurring words in our tweets dataset. We also created a frequency distribution chart representing the top 30 most frequent words on the x-axis and their frequency on the y-axis. Refer to figure 2.

(4) **Classification through various models and finding their accuracies:** Considering the classification output of the Random-Forest model as the baseline result, we splitted the output into 20-80% and used it to train and test on two other classification models: Naive Bayes and XGB classifier.

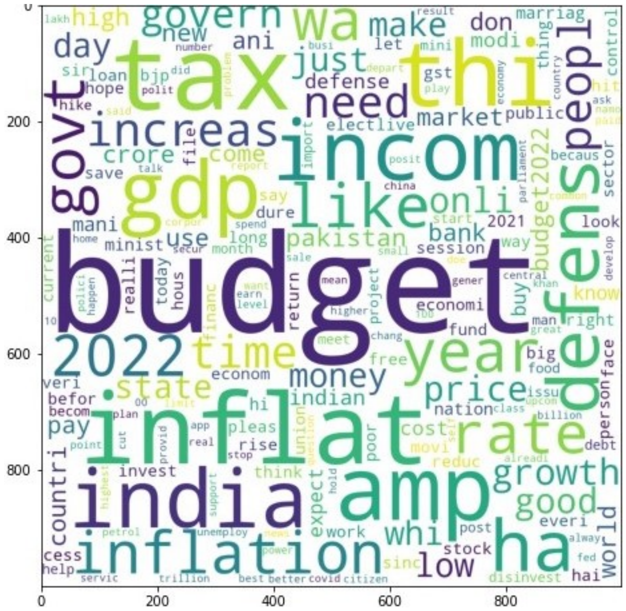

Figure 2: Word cloud for pre budget phase

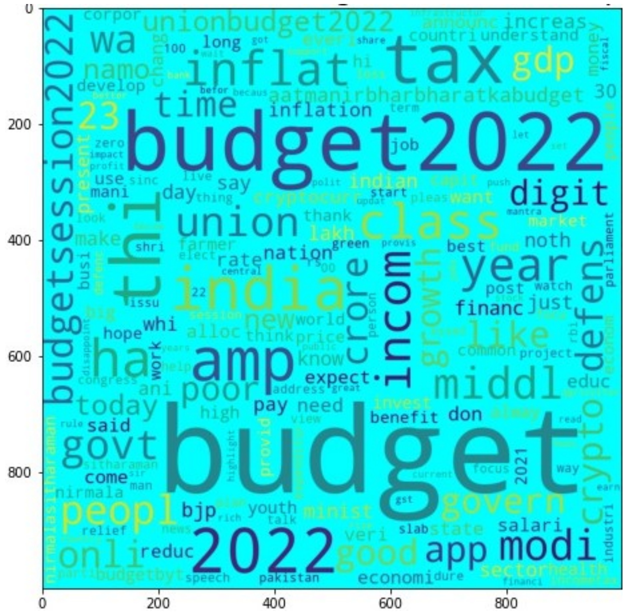

Figure 3: Word cloud for post budget phase

- **Post-budget phase analysis:** Similar to Pre-budget phase, sentiment analysis, emotional analysis, visualization using word cloud and frequency distribution and classification through other models was done and calculate corresponding accuracies. Refer to figure 3.

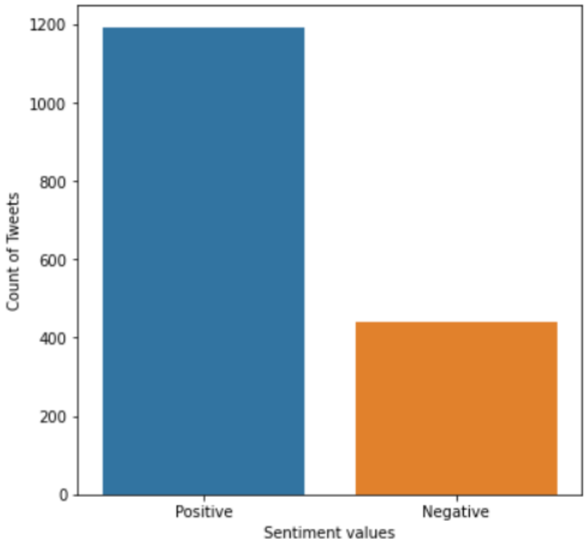

Figure 4: Public Opinion on Union Budget 2022 for pre-budget phase

7 BASELINE RESULTS

- **Pre-budget Phase:** On performing Sentiment analysis on the tweets prior to the Budget, there were approximately 1200 'Positive' tweets and 400 'Negative' tweets.
  Emotional classification using the Random-Forest model resulted in approximately 1200 tweets that were labeled as 'joy',190 tweets as 'neutral', 150 as 'sadness' followed by 'fear', 'anger', and 'surprise' that were less than 100.However, tweets with labeling as 'shame' and 'disgust' were negligible. Refer to figure 4
  The Word Cloud and the Frequency distribution chart for the top 30 most occurring words showed that the words like 'budget', 'tax', 'inflation', 'income' 'defense' were most mentioned in the tweets.
  On considering the above classification result as a baseline, splitting and using it for training and testing a Multinomial Naive Bayes classifier model, the results gave an accuracy of 78%. However, using the XGB classifier resulted in an accuracy of 83%. Refer to figure 5.
- **Post-budget Phase:** On performing sentiment analysis on tweets and comments replies from the youtube videos post the Budget, there were approximately 3500 'Positive' tweets and 1000 'Negative' tweets.
  Doing emotion classification using the Random forest model resulted in approximately 500 as 'neutral', 3000 as 'joy', 490 as 'sadness', 50 as 'fear', 100 as 'surprise',120 as 'anger', and around 10 as 'disgust' emotions. Refer to figure 6.
  On considering the above classification result as a baseline, splitting and using it for training and testing a Multinomial Naive Bayes classifier model, the results gave an accuracy of 74%. However, using the XGB classifier resulted in an accuracy of 87%. Refer to figure 7.

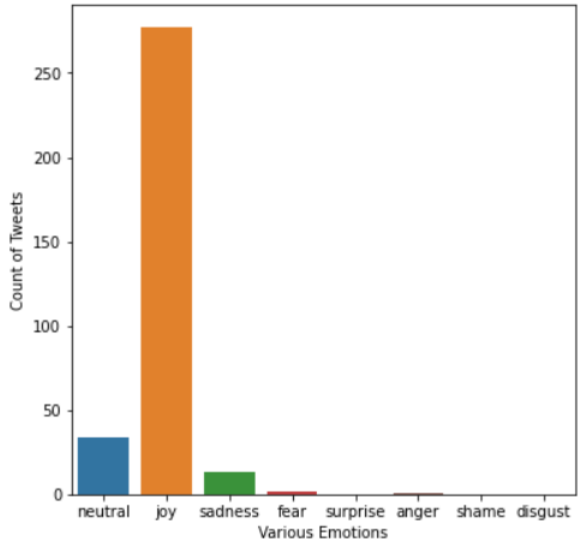

Figure 5: Pre-Budget analysis through XGB classifier

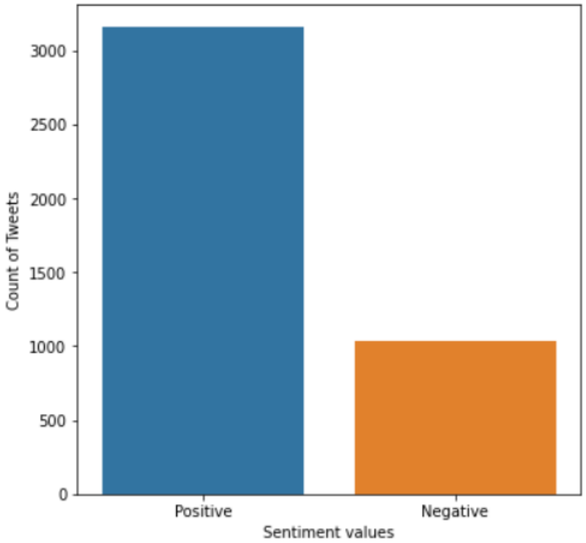

Figure 6: Public Opinion on Union Budget 2022 for post-budget phase

8 NOVELTY

Our work is different from others in the sense that:

- Dataset corpus is made using social media platforms like twitter and youtube.
- Classifying tweets in different sets of emotions for example joy, anger, sadness, trust, etc. Along with that sentiments were grouped into 3 categories i.e. positive, neutral, and negative.
- Try to compare the tweets of budget 2021-22 with 2022-23 and try to analyze the difference that has come in Indian society in a gap of 1 year.

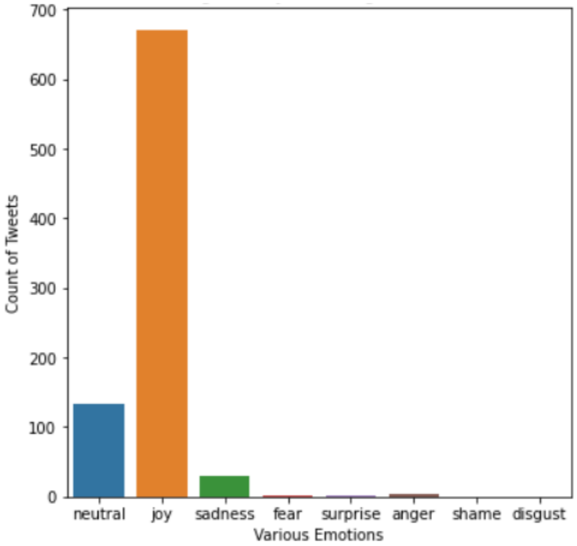

Figure 7: Post-Budget analysis through XGB classifier

- Analysis performed in 2 phases i.e. Pre-phase(before release of budget) and post-phase(after release of budget).

9 FUTURE SCOPE

- Compare tweets of more than 1 previous year based on data availability.
- Analyze the tweets tweeted in multiple language across the country.
- Sentimental Analysis is done after classifying the tweets into major sectors of budget i.e. finance, education, agriculture, service, defense, etc, and correspondingly analyzing the budget impact on the above-mentioned sectors.

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. https://doi.org/10.21203/rs.3.rs-1515523/v1

[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. https://doi.org/10.18653/v1/S15-2098

[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 Singh3 Rupinder Kaur1, Rajvir Kaur2 and Dr. Sandeep Ranjan4. 2020. Twitter Sentiment Analysis of the Indian Union Budget 2020. 29 (2020), 8. http://www.ijsrcsams.com/images/stories/Past_Issue_Docs/ijsrcsamsv7i4p149.pdf

[6] Moonis Shakeel and Vikram Karwal. 2016. Lexicon-based sentiment analysis of Indian Union Budget 2016–17. (july 2016), 299–302. https://doi.org/10.1109/ICSPCom.2016.7980595

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