

# Analysis of Times of India Headlines in 2023

Avanish Mishra, Gopalakrishnan M, Hrishant Singhal, Ranu Patel

October 7, 2024

# Motivation

Reading Newspapers daily have become an eternal part of many people and headlines generally are flashy to attract more audience and make it catchy. We are trying to look whether these headlines try to convey the purpose or is it misleading (bogus). Through various text analysis methods and libraries, we have tried to explore varies domains that are issues of day-to-day like politics, environment etc.



# Sentiment of Major News

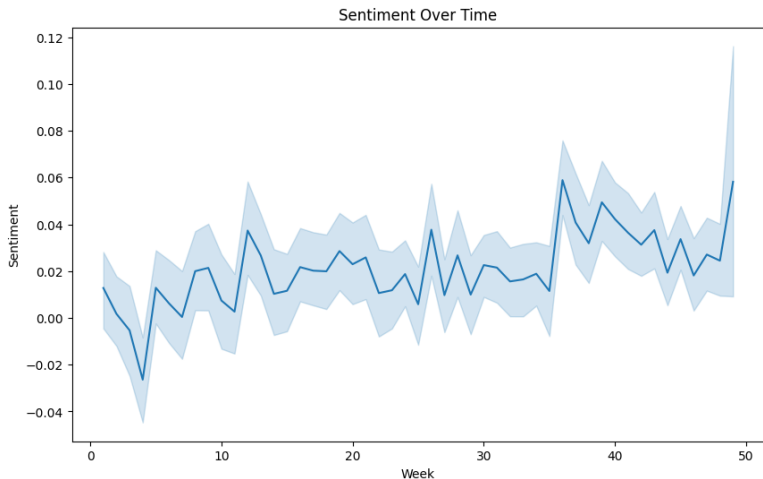


Figure: Sentiment of major news

# Analysis of Political News



Figure: Word Cloud for Political News



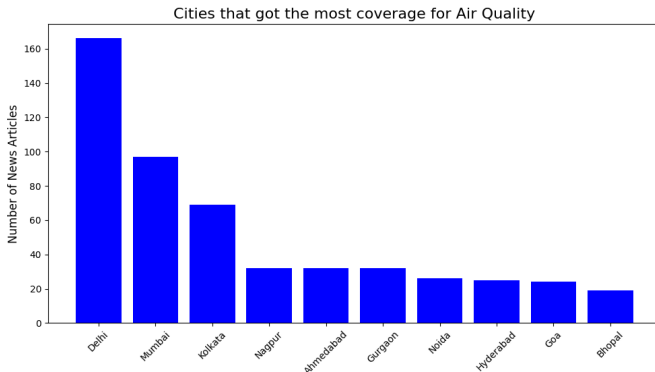






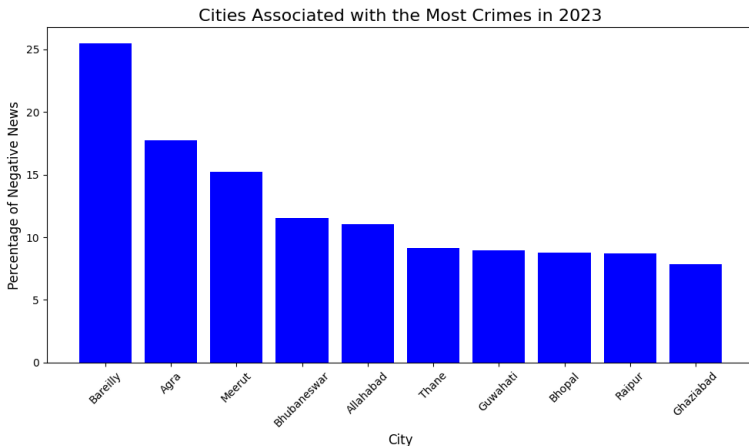
# Analysis on Environmental issues

- Firstly, we filtered out headlines which had the keywords such as pollution and air quality.
- Next, from the links we took the cities which the filtered headlines were reported from.
- We could find that Delhi has the highest. Cities like Bangalore, Ahmedabad were also having high pollution but weren't mentioned in the news with high frequency.



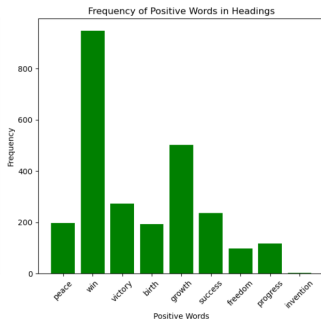
# Crime Analysis Across Cities

- We first filtered headlines containing any of the following words: ['murder', 'kill', 'rape', 'violence', 'murdered', 'killed', 'raped'].
- Then we identified the name of the cities from the link of the webpage of that headline.



# Positive vs Negative Sentiments

- Negative Words Dominance: Words like 'murder', 'rape', and 'violence' appear most frequently, indicating a focus on sensational or shocking events in headlines.
- Positive Words Clustered: Positive news, focusing on achievements like 'win' and 'growth', is covered but less frequent than negative news, indicating less diverse coverage of positive events.



# Analysis on selection of topics

- For finding the topics, we understood that there is a scikit learn tool (Github) called Term Frequency inverse document vectorizer which helps in finding the number of times terms have occurred and giving vector form to the arrangement of words and Non-negative Matrix Factorization (NMF) for identifying collection of latent topics.
- This includes defining the vectorizer, fitting the model, using a loop for selecting the words.
- When we looked at it for the headlines, we could find that the major topics were around politics, crimes etc.
- For Example, Topic 5 has high court, order, plea, seek; these are words which are interconnect and comes under a major topic of judiciary.

```
Topic 1:  
bjp, govt, cm, delhi, state  
Topic 2:  
man, arrested, wife, death, minor  
Topic 3:  
case, covid, murder, tamil, nadu  
Topic 4:  
woman, dy, cop, killed, accident  
Topic 5:  
hc, court, order, plea, seek
```