

# Text Analytics

## Business Insight Report on Popular Newspapers:

- Wall Street Journal
- Times of India
- Hindustan Times
- The Indian Express

## Sentiment Analysis of Top 20 Trending Articles on Indian National News

19 MARCH 2020

Hult International Business School

Authored by: Arundhishaan Kanagaraja

Cohort: MSBA5 - Valencia



---

# Introduction

English-language dailies and journals remain highly influential in India. The vernacular press is increasing steadily in absolute and relative importance. Overall, the press functions with little government censorship, serious controls in matters of national security, in times of emergency or to avoid inflaming passions that lead to communal riots or comparable disturbances (Britannica, 2019).

***“The largest-circulating dailies are The Times of India and Hindustan Times (both in English), the Hindustan and the Navbharat Times (Hindi), and the Anandabazar Patrika (Bengali).”***

This proves a great opportunity of unstructured text data that is available for data analysis. This report focuses on the Sentiment Analysis of Textual data from four popular Newspapers: Wall Street Journal (WSJ), Times of India (TOI), Hindustan Times and The Indian Express. The latter three are top English newspapers in India.

## *Framework 1*

Tokenization and bi-grams are used to initially understand the most common news by the newspaper. Further, the sentiment analysis is done using the ‘nrc’ dictionary in R. We would visualize the results using bar charts and word clouds.

## *Framework 2*

Term Frequency – Inverse Document Frequency (tf-idf) is used to find the unique words from articles of these Newspapers that gives us the insight of how these Newspapers are different in their focus to pick specific News.

---

# Data Collection

Trending top 20 Online News articles are collected from the Indian National News section of each newspaper's website which is saved in pdf format. Thus, we have obtained our unstructured data.

## Data Cleaning

The pdf documents of 20 articles are imported and structured into a dataframe. Unwanted text such as ads, author names, non-insightful repetitive text, etc. are eliminated by defining custom stopwords. The column name defines the newspaper company that the articles are taken from and the rows contain the text data from the articles.

Similar dataframes are created for the remaining three newspapers.

## Text Tokenization & Bi-grams

The following are the tidy format of tokenized data of different newspapers:

### ➤ *Wall Street Journal*

```
# A tibble: 3,004 x 2
  word          n
  <chr>      <int>
1 law          195
2 citizenship  177
3 india's      165
4 government   148
5 muslims      140
6 muslim       125
7 protests     123
8 delhi        120
9 police       117
10 hindu        95
# ... with 2,994 more rows
```

```
# A tibble: 3,004 x 3
  word1 word2          n
  <chr>  <chr>      <int>
1 citizenship law        72
2 prime  minister    54
3 minister narendra    42
4 muslim  majority    42
5 hindu   nationalist   33
6 bharatiya janata     30
7 janata  party         28
8 rights  reserved        28
9 citizenship amendment  26
10 internet services     22
# ... with 2,994 more rows
```

**Insights:** The most common news by Wall Street Journal follows information related to Law, actions by the Indian government, social protests and views of the people. From the bi-gram, we can further understand that there is news widely spoken about citizenship amendments, people rights and the actions of the ruling party.

### ➤ *Times of India*

# A tibble: 1,946 x 2	# A tibble: 1,788 x 3
word n	word1 word2 n
<chr> <int>	<chr> <chr> <int>
1 delhi 573	1 delhi election 110
2 election 326	2 election result 104
3 2020 193	3 delhi assembly 100
4 assembly 132	4 assembly election 59
5 congress 129	5 delhi delhi 59
6 result 106	6 delhi polls 59
7 aap 99	7 2020 liveblog 58
8 bjp 99	8 congress candidates 52
9 party 96	9 arvind kejriwal 51
10 polls 91	10 63 congress 49
# ... with 1,936 more rows	# ... with 1,778 more rows

**Insights:** The most common news published by Times of India covers information about the 2020 Delhi elections. From the bi-gram, we can further understand that there is news widely spoken about the victory of Arvind Kejriwal as the Chief Minister of Delhi (Dutt, 2020).

### ➤ *Hindustan Times*

# A tibble: 1,896 x 2	# A tibble: 1,550 x 3
word n	word1 word2 n
<chr> <int>	<chr> <chr> <int>
1 court 46	1 board exams 32
2 election 34	2 omar's detention 21
3 board 33	3 chief subhash 20
4 cricket 32	4 cong chief 20
5 exams 32	5 crushing defeat 20
6 school 32	6 party's crushing 20
7 sports 32	7 1st casualty 19
8 party 29	8 defeat psa 19
9 people 28	9 hear challenge 19
10 chief 27	10 kejriwal's party 19
# ... with 1,886 more rows	# ... with 1,540 more rows

**Insights:** The most common news by Hindustan Times follows information related to Politics, Education and Sports. From the bi-gram, we can further



understand that there is news widely spoken about higher secondary Board Examinations, elections and specific political leaders.

### ➤ *The Indian Express*

# A tibble: 1,692 x 2	# A tibble: 1,327 x 3
word n	word1 word2 n
<chr> <int>	<chr> <chr> <int>
1 delhi 34	1 honey mahajan 8
2 police 31	2 rear admiral 8
3 government 28	3 worth rs 8
4 court 20	4 assembly polls 7
5 minister 18	5 chief minister 6
6 rs 18	6 farm house 6
7 women 18	7 government hospitals 6
8 bjp 16	8 mobile phones 6
9 party 16	9 sanjay vatsayan 6
10 people 16	10 admiral sanjay 5
# ... with 1,682 more rows	# ... with 1,317 more rows

**Insights:** The most common news by The Indian Express follows information related to Politics and Law. From the bi-gram, we can further understand that there is also news widely spoken about Military Personnel ('Admiral' – Commander of Naval squadron), Technology and Commerce.

## Sentiment Analysis & Word Clouds

The following are the Sentiment Analysis in tidy format, bar chart & wordcloud (using nrc dictionary):

### ➤ *Wall Street Journal*

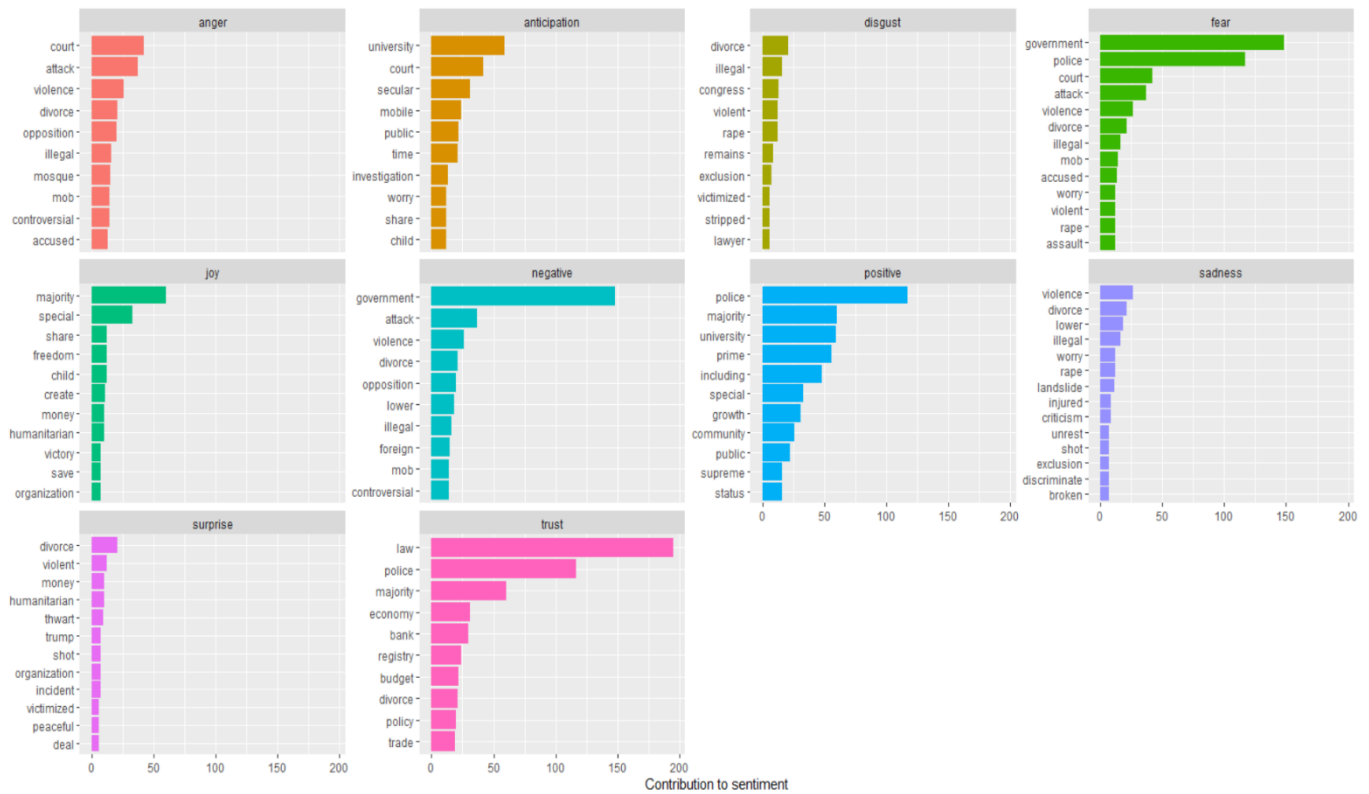
# A tibble: 344 x 3	# A tibble: 333 x 2	# A tibble: 1,409 x 3
word sentiment n	word n	word sentiment n
<chr> <chr> <int>	<chr> <int>	<chr> <chr> <int>
1 protests negative 123	1 protests 123	1 law trust 195
2 critics negative 46	2 critics 46	2 government fear 148
3 attack negative 37	3 attack 37	3 government negative 148
4 protest negative 35	4 protest 35	4 police fear 117
5 support positive 22	5 growth 31	5 police positive 117
6 opposition negative 20	6 violence 26	6 police trust 117
7 illegal negative 16	7 protesters 24	7 majority joy 60
8 supreme positive 16	8 arrested 23	8 majority positive 60
9 controversial negative 14	9 banned 22	9 majority trust 60
10 led positive 14	10 support 22	10 university anticipation 59
# ... with 334 more rows	# ... with 323 more rows	# ... with 1,399 more rows

- **bing dictionary**

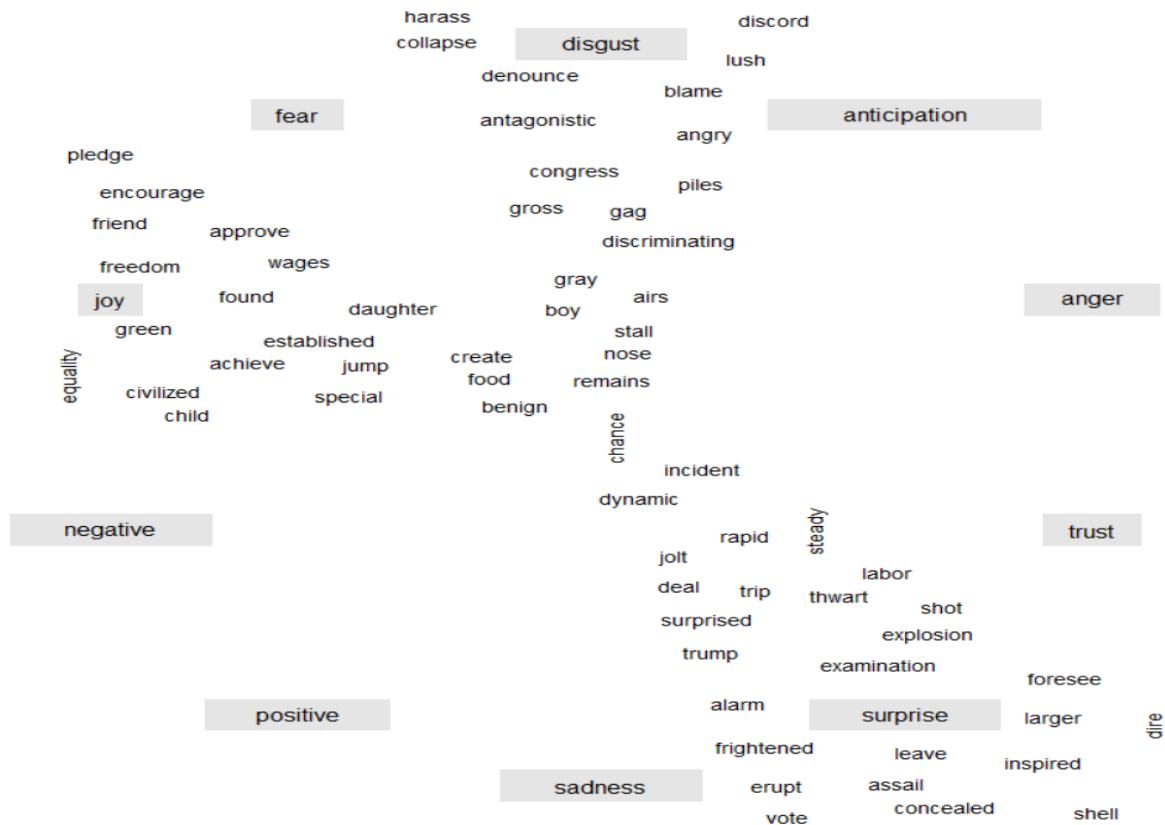
- **afinn dictionary**

- **nrc dictionary**

## Bar Chart (Using nrc dictionary showing words' contribution to sentiment):



## Word Cloud (Using nrc dictionary):



**Insights:** From the above charts, we can understand that Wall Street Journal articles have a strong use of words that relate to the sentiments, surprise and joy. When they publish political News, they also focus more on the views of the people which can be seen through the usage of the word ‘protest’ and other words used closer to the ‘disgust’ sentiment in the word cloud.

### ➤ Times of India

```
# A tibble: 170 x 3
  word      sentiment      n
<chr>    <chr>    <int>
1 lose      negative    71
2 trump     positive    65
3 victory   positive    63
4 fans      positive    60
5 tortured  negative    60
6 lost      negative    50
7 top       positive    48
8 breaking  negative    40
9 excitement positive    40
10 losing    negative    36
# ... with 160 more rows
```

- **bing dictionary**

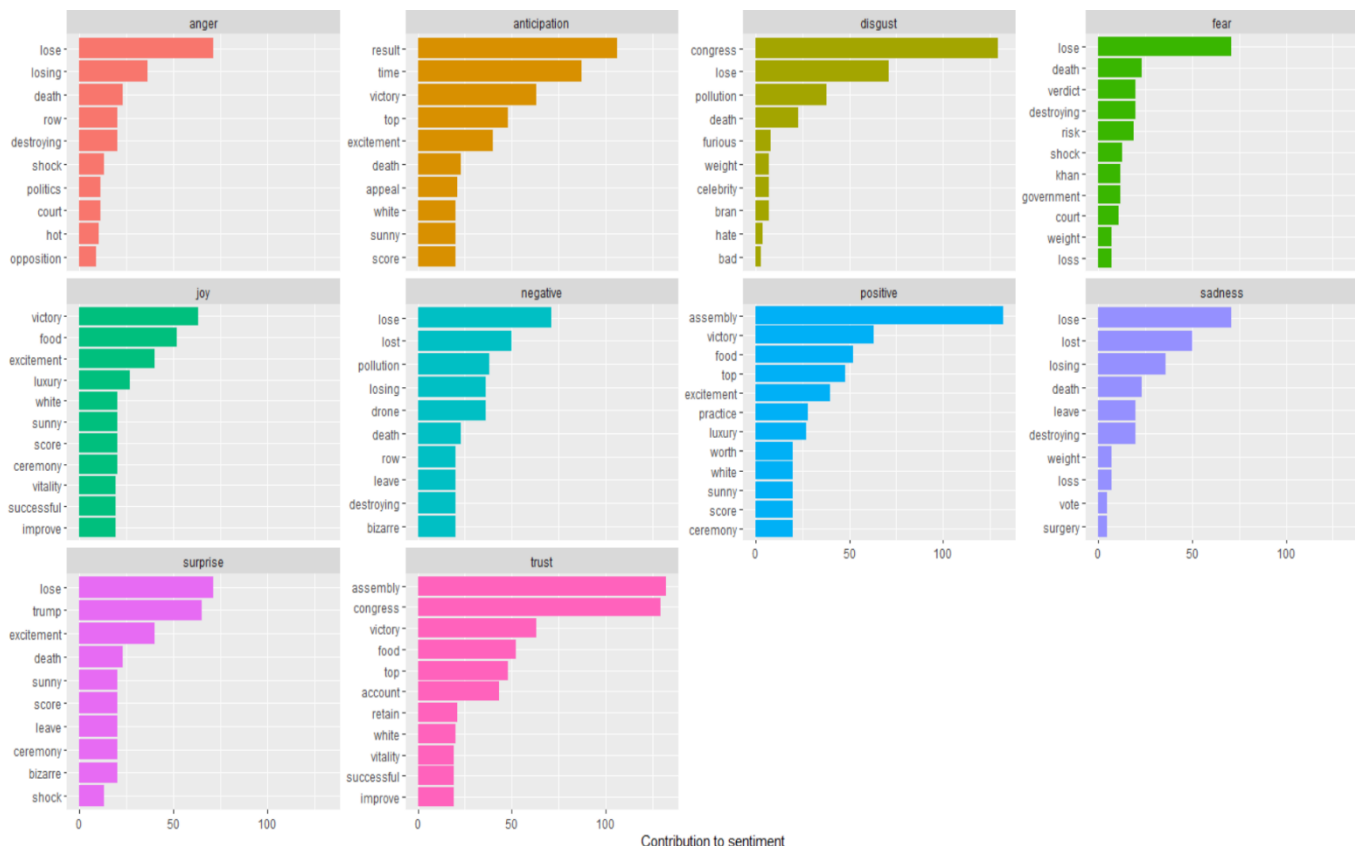
```
# A tibble: 150 x 2
  word      n
<chr>    <int>
1 stop      63
2 tortured  60
3 lost      50
4 top       48
5 excitement 40
6 shares    40
7 losing    36
8 treasures 30
9 free      29
10 popular  26
# ... with 140 more rows
```

- **afinn dictionary**

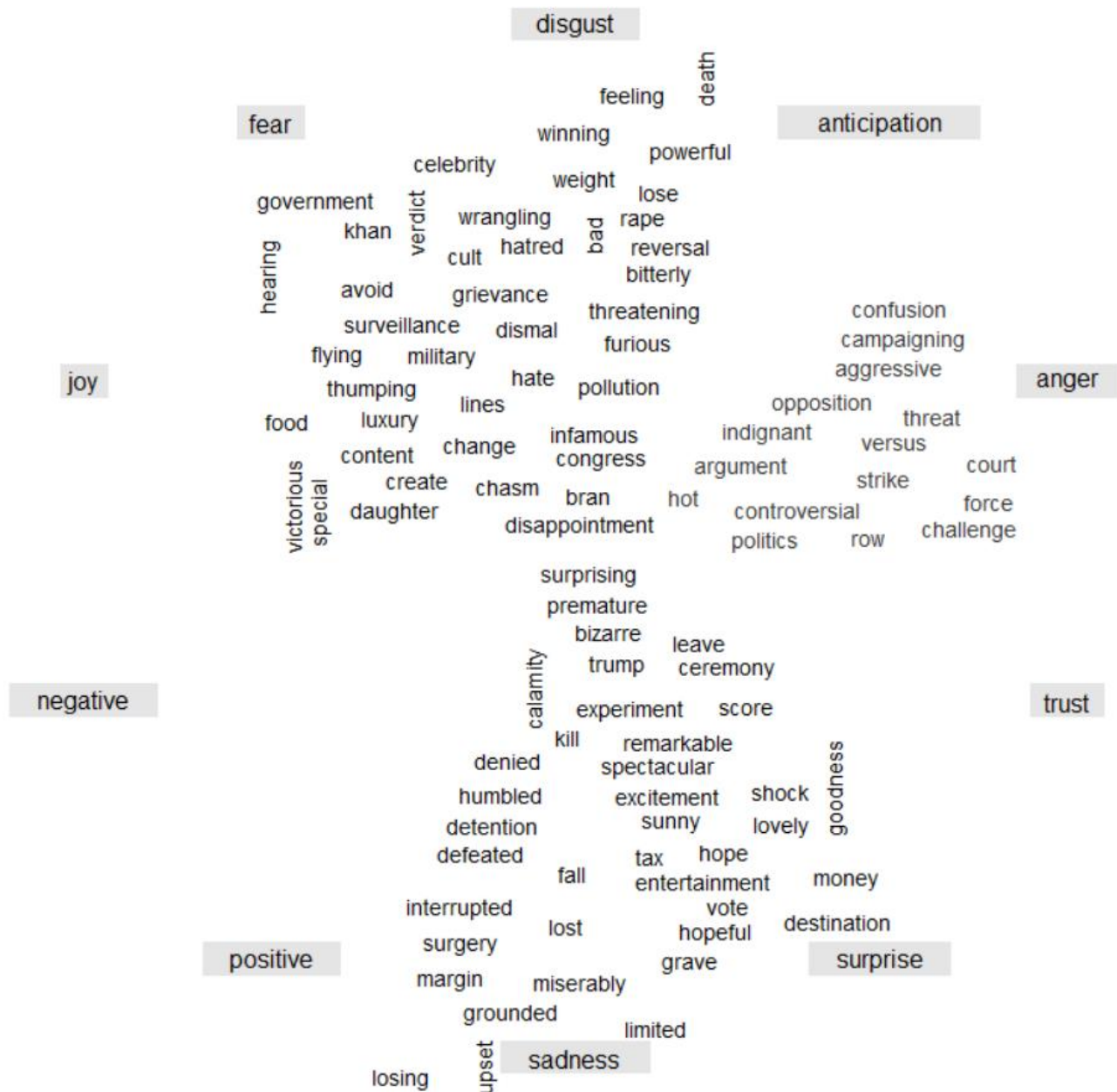
```
# A tibble: 662 x 3
  word      sentiment      n
<chr>    <chr>    <int>
1 assembly positive    132
2 assembly trust      132
3 congress disgust    129
4 congress trust      129
5 result  anticipation  106
6 time    anticipation   87
7 lose    anger          71
8 lose    disgust        71
9 lose    fear          71
10 lose    negative        71
# ... with 652 more rows
```

- **nrc dictionary**

Bar Chart (Using nrc dictionary showing words' contribution to sentiment):



## Word Cloud (Using nrc dictionary):



**Insights:** From the above charts, we can understand that Times of India articles have a strong use of words that relate to the sentiments, fear, disgust, sadness and anger. The word 'lose' has been used to drive different sentiments which is related to covering the Delhi Elections. The words 'assembly' and 'congress' also indicate that they are publishing many stories about political parties actively.



## ➤ Hindustan Times

# A tibble: 189 x 3

	word	sentiment	n
	<chr>	<chr>	<int>
1	defeat	positive	24
2	crushing	negative	20
3	victory	positive	20
4	wins	positive	20
5	casualty	negative	19
6	limited	negative	19
7	promises	positive	19
8	fall	negative	18
9	shine	positive	16
10	rape	negative	11

# ... with 179 more rows

- **bing dictionary**

# A tibble: 170 x 2

	word	n
	<chr>	<int>
1	detention	26
2	challenge	22
3	crushing	20
4	wins	20
5	casualty	19
6	limited	19
7	promises	19
8	rape	11
9	death	9
10	accused	8

# ... with 160 more rows

- **afinn dictionary**

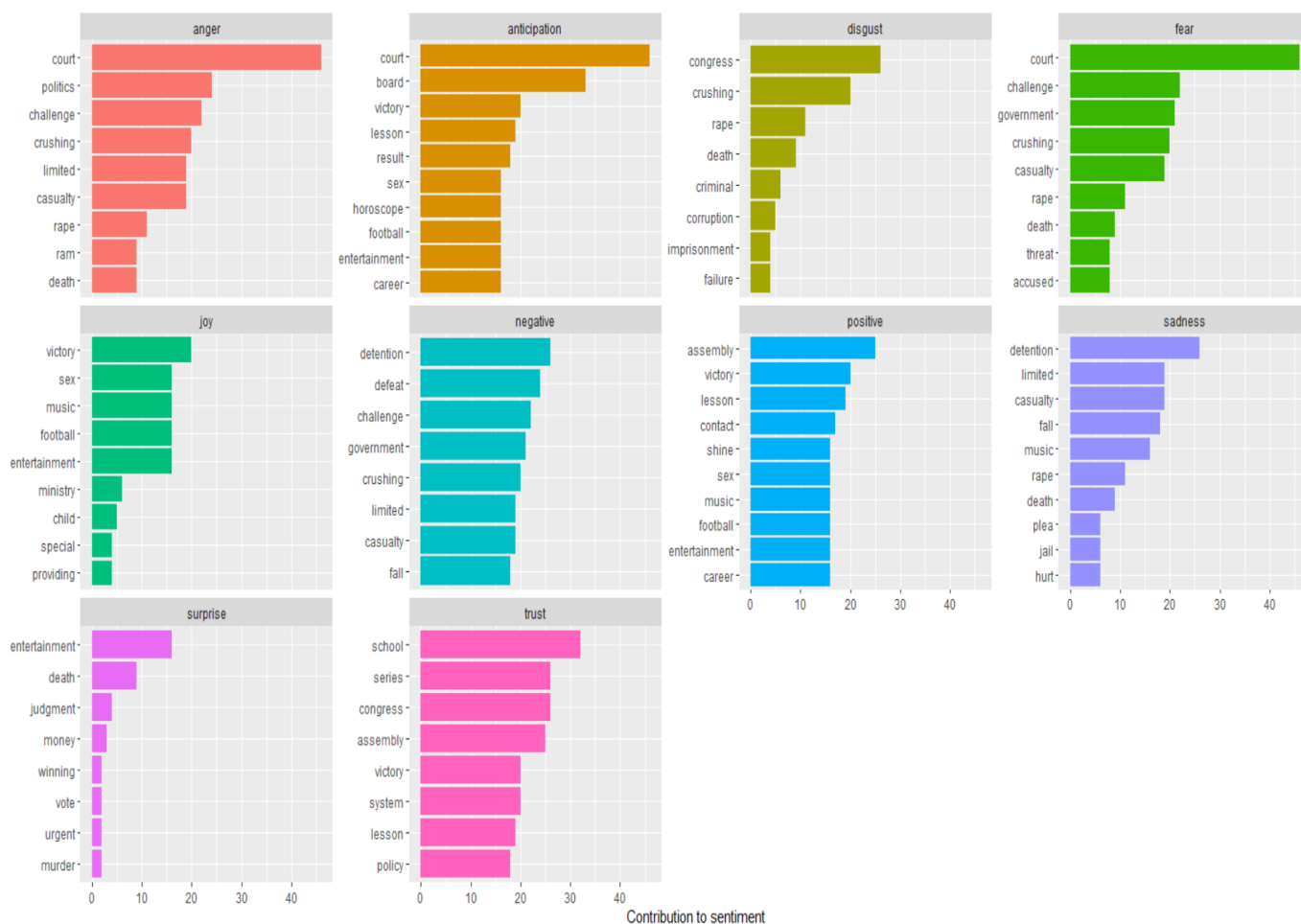
# A tibble: 834 x 3

	word	sentiment	n
	<chr>	<chr>	<int>
1	court	anger	46
2	court	anticipation	46
3	court	fear	46
4	board	anticipation	33
5	school	trust	32
6	congress	disgust	26
7	congress	trust	26
8	detention	negative	26
9	detention	sadness	26
10	series	trust	26

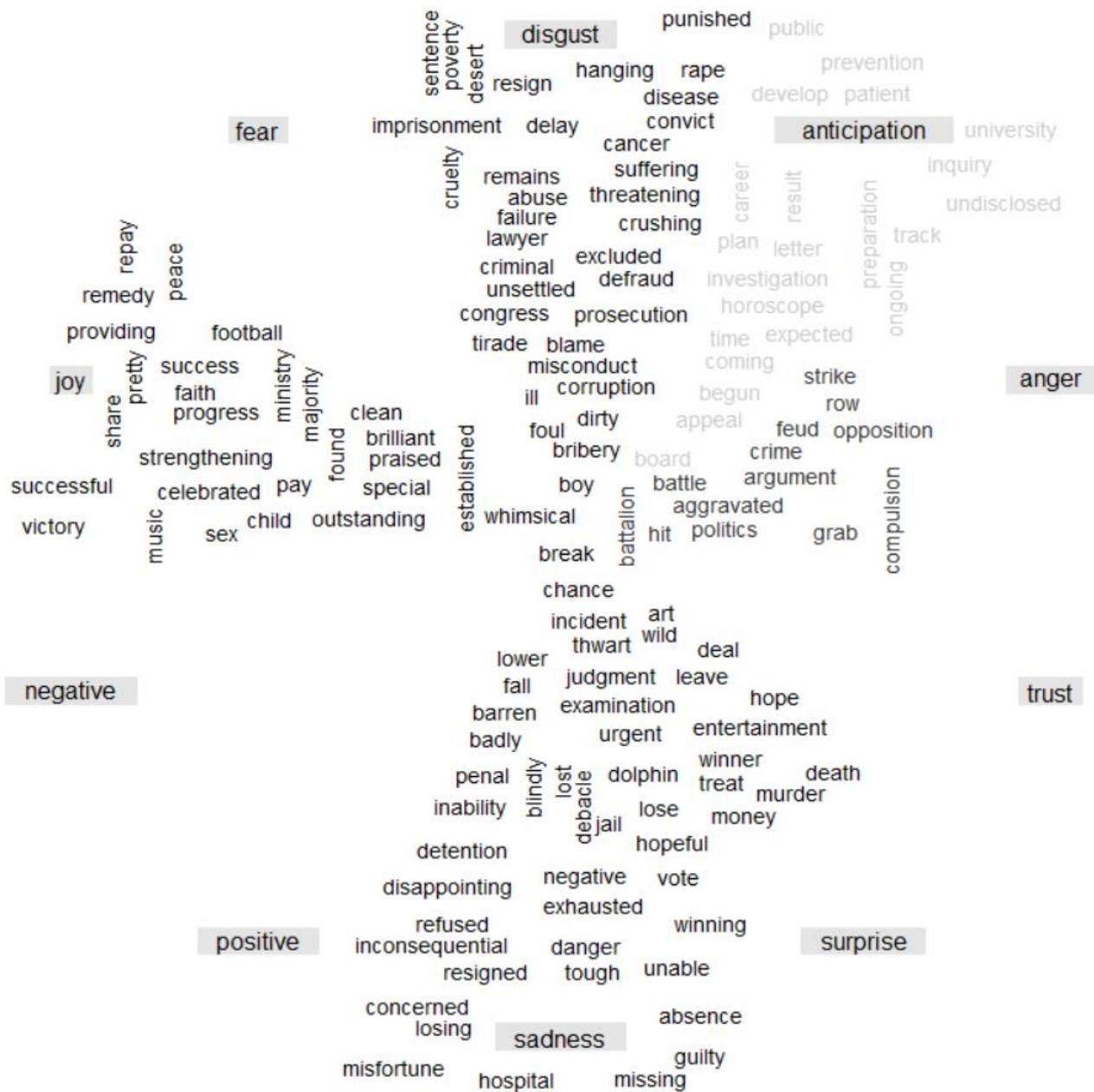
# ... with 824 more rows

- **nrc dictionary**

Bar Chart (Using nrc dictionary showing words' contribution to sentiment):



### Word Cloud (Using nrc dictionary):



**Insights:** From the above charts, we can understand that Hindustan Times articles have a strong use of words that relate to the sentiments, anticipation, sadness, disgust and joy. The word 'court' has been used to drive different sentiments like anger, anticipation and fear. This relates to News on Law cases where readers anticipate the story. Similar words like 'criminal', 'punished', 'rape', etc. support the same sentiments.

## ➤ The Indian Express

# A tibble: 161 x 3

	word	sentiment	n
	<chr>	<chr>	<int>
1	attack	negative	8
2	protest	negative	8
3	worth	positive	8
4	protesting	negative	7
5	appeal	positive	5
6	rejected	negative	5
7	forged	negative	4
8	mercy	positive	4
9	plea	negative	4
10	rape	negative	4

# ... with 151 more rows

- **bing dictionary**

# A tibble: 133 x 2

	word	n
	<chr>	<int>
1	join	15
2	attack	8
3	protest	8
4	worth	8
5	justice	7
6	protesting	7
7	accused	5
8	rejected	5
9	arrested	4
10	mercy	4

# ... with 123 more rows

- **afinn dictionary**

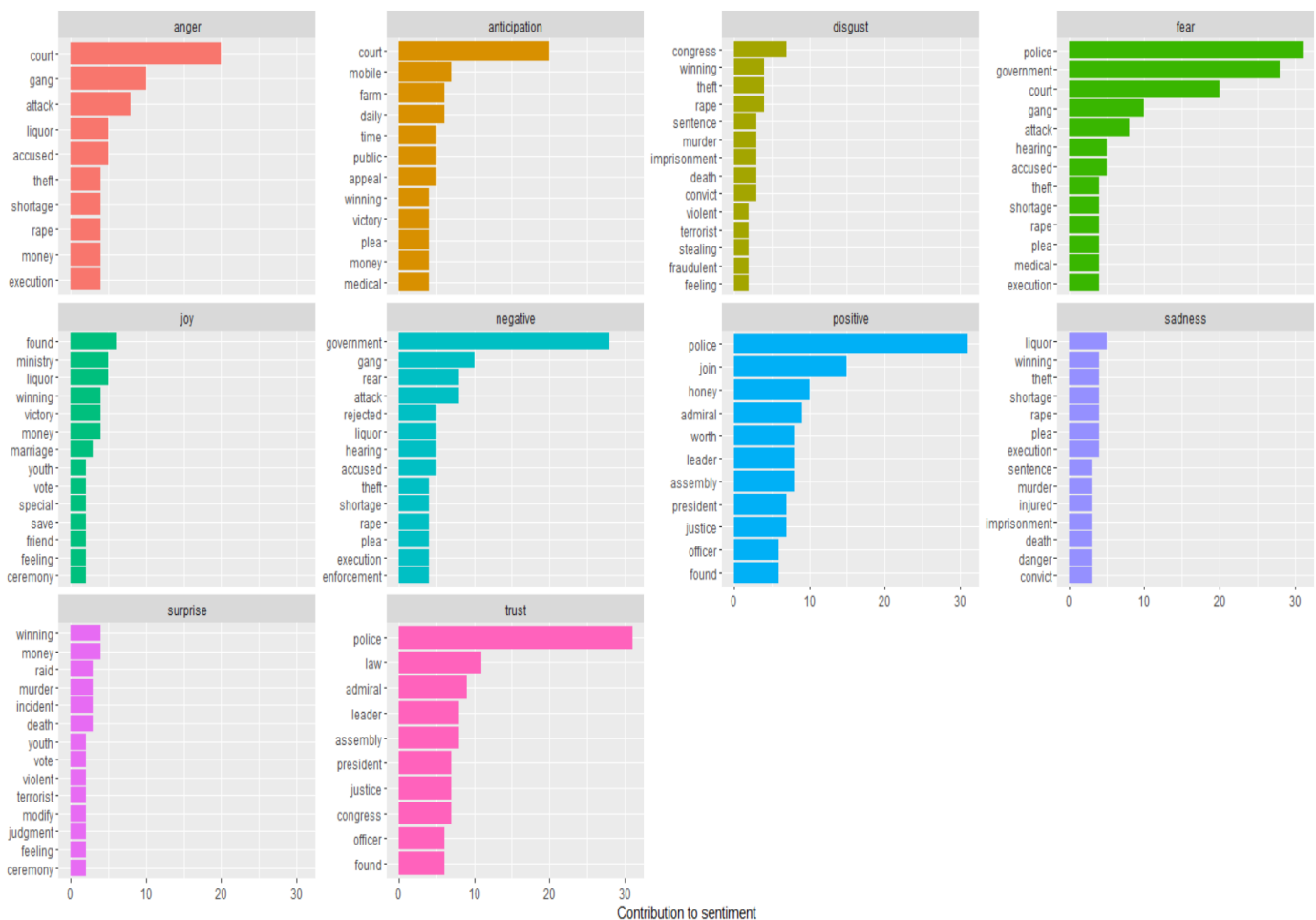
# A tibble: 706 x 3

	word	sentiment	n
	<chr>	<chr>	<int>
1	police	fear	31
2	police	positive	31
3	police	trust	31
4	government	fear	28
5	government	negative	28
6	court	anger	20
7	court	anticipation	20
8	court	fear	20
9	join	positive	15
10	law	trust	11

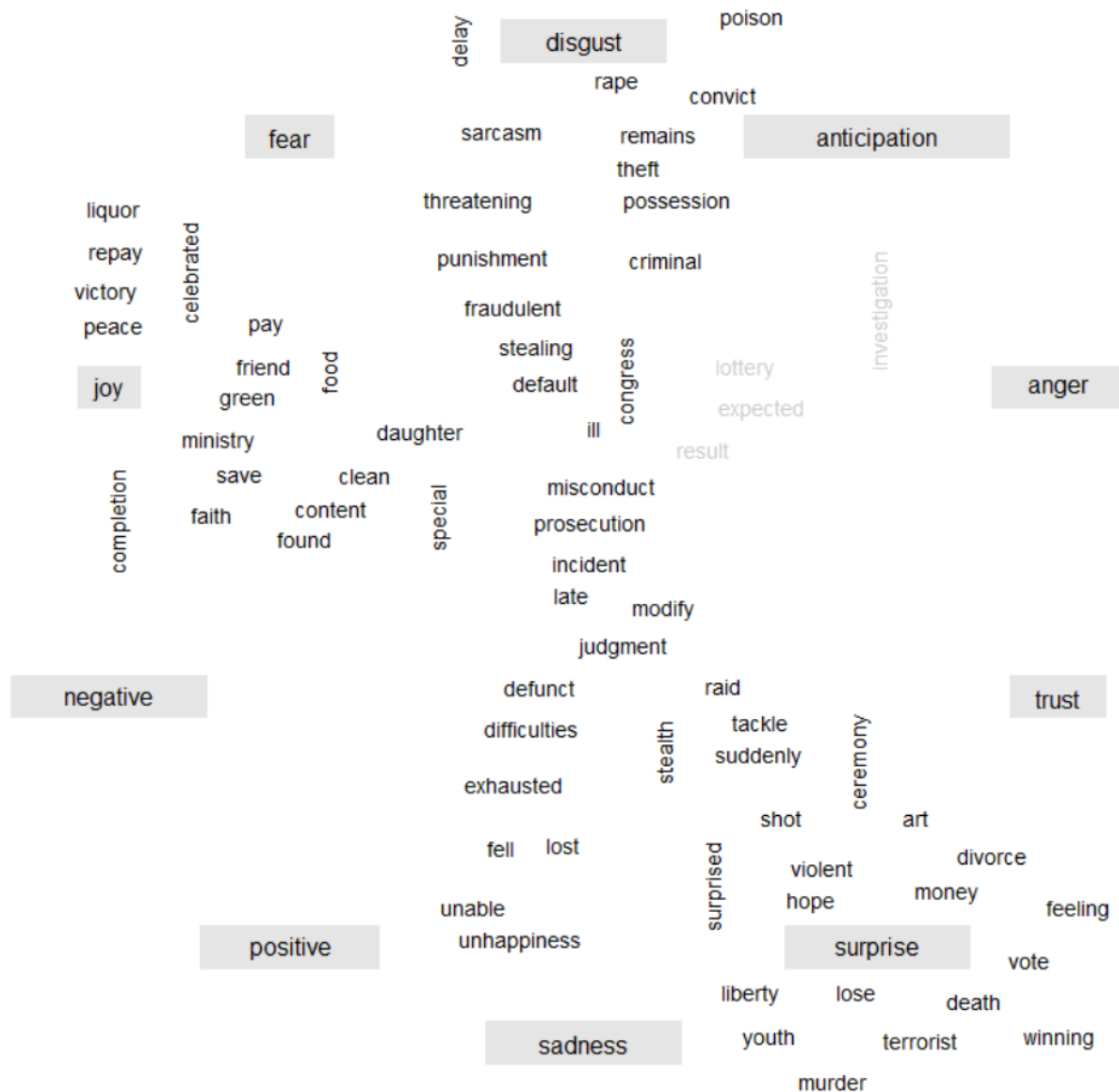
# ... with 696 more rows

- **nrc dictionary**

Bar Chart (Using nrc dictionary showing words' contribution to sentiment):



### Word Cloud (Using nrc dictionary):

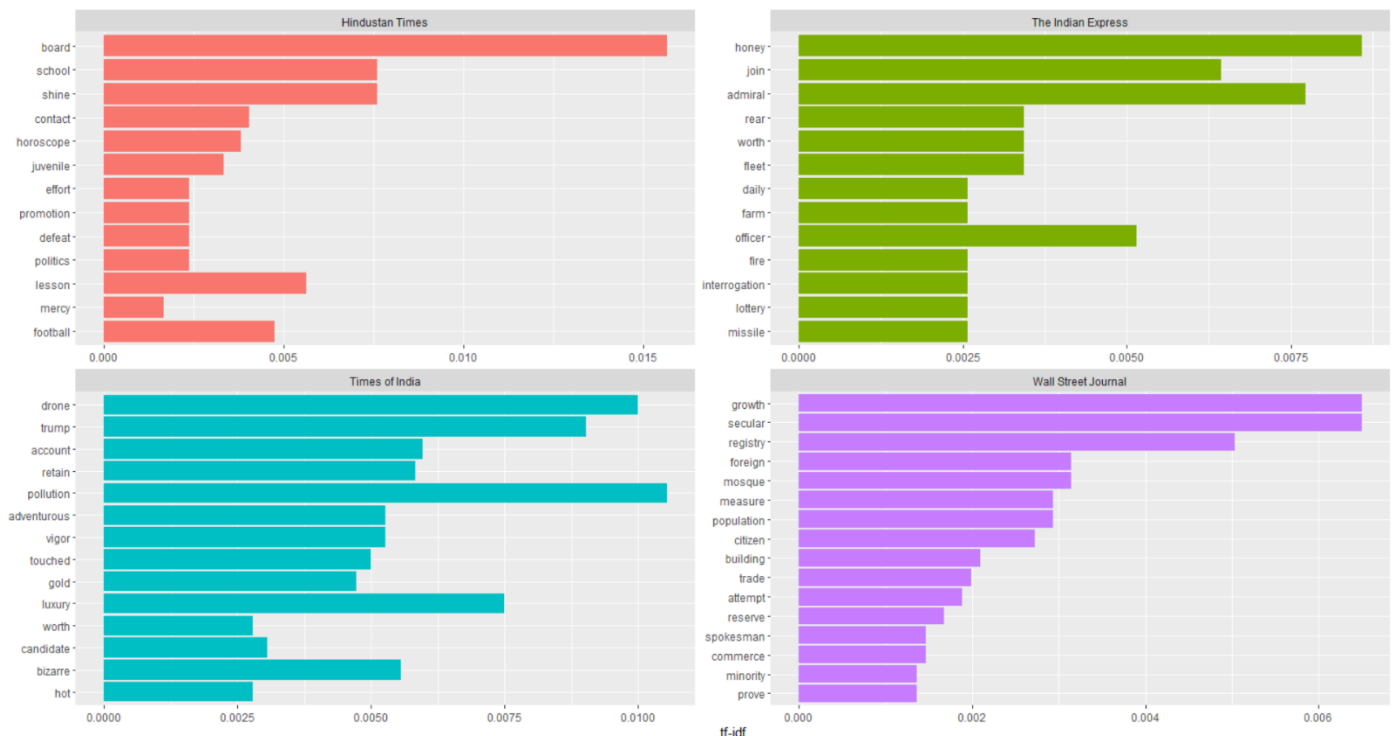


**Insights:** From the above charts, we can understand that The Indian Express articles have a strong use of words that relate to the sentiments, surprise, disgust and joy. The word 'police', 'government' and 'court' are widely used which contribute to different sentiments which involves News on several Law cases. The words 'money', and 'winning' contribute to the sentiment surprise, and joy indicating commerce or social News.

# Finding Unique Trending News

The term frequency - inverse document frequency (tf-idf) is used to identify unique words from the articles of each newspaper. Though there are very common trending News, using tf-idf we can identify how newspapers differentiate themselves to their customers.

Bar graph using tf-idf:



## Insights:

- Wall Street Journal focuses on a very broader scale of News comprising growth, people's rights, culture, trade and commerce.
- Times of India focuses on luxury, gold, technology and environmental News targeting youth, women and activists.
- Hindustan Times focuses on education and sports news targeting youth and entertainment sector.
- The Indian Express focuses more on specific Law cases, military news and other anticipatory news like lottery, social news, etc.



---

# Conclusion

By using Text analytics techniques, unstructured trending News data on Indian National News of different Newspapers is transformed into structural data for data analysis.

The Sentiment analysis shows that different Newspapers express various sentiments using different style of writing to sell their News to their customers.

- Wall Street Journal sets a standard perspective to share quality information to students and professional class of people.
- Newspapers may focus on every aspect of a specific topic. For example, Times of India focuses on covering the Delhi Elections.
- Newspapers may also create awareness on pollution, new technology, education, etc.
- Newspapers may focus on articles that create surprise, joy and anticipation. For example, The Indian Express focuses on Law cases, social events, lottery, etc.

Finally, we found some unique news, sectors and customers that each Newspaper is targeting.

---

# Reference

The Editors of Encyclopaedia Britannica. (2019, February 4). India. Retrieved from <https://www.britannica.com/place/India/Cultural-institutions#ref487381>

Dutt, B. (2020, February 11). Opinion | How this Delhi leader fought and defeated Modi's BJP - and why it matters. Retrieved from <https://www.washingtonpost.com/opinions/2020/02/11/how-this-delhi-leader-fought-defeated-modis-bjp-why-it-matters/>