SIL801
Project Report

Influence of celebrity media choices on their followers

Authors

Nikhil Verma

Nipun Gupta

Sachin Yadav

Supervisor

Prof. Aaditeshwar Seth

Contributor

Amit Ruhela

Introduction

Studying factors which contribute towards information bias is a recurring theme we have encountered in the course. In this project we set out to analyse the influence that the media choices of celebrities have on those of their followers. For this we collect a large Twitter dataset totalling over 450 million tweets and process it so we finally have a distribution of celebrities' and their followers' engagement factors for each of the media houses defined in mediaDomains, across the categories of Bollywood, Sports and Politics. Finally we do statistical analysis to ascertain the correlation between these two. Our initial hypothesis is that since subjectivity in news is present in higher degree in political content as compared to factual reporting of sports and entertainment events, the influence should be higher in Politics than in Sports and Bollywood.

Research Question

- *Do the media choices of a small minority of very popular users on Twitter influences their followers' media choices, who form the bulk population of ordinary users ?*

- *If so, how does this vary across different classes of celebrities, namely Entertainment , Sports & Politics ?*

Code Repositories

UrlExpander	https://github.com/nikhilaii93/UrlExpander
TweetsAnalyzer	https://github.com/nikhilaii93/TweetsAnalyzer

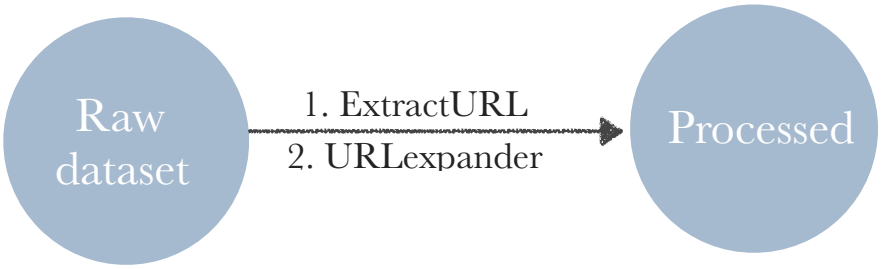
Original Dataset

The original dataset , with core information extracted to 120GB , consists of Tweets over a period of 95 days.
A snapshot of the extracted tweet database:

Dataset	Seed users	Followers (millions)	Tweets (millions)
Entertainment	150	23	406
Politics	55	7	115
Sports	40	9	129
Total	**245**	**26**	**468**

Processed Database

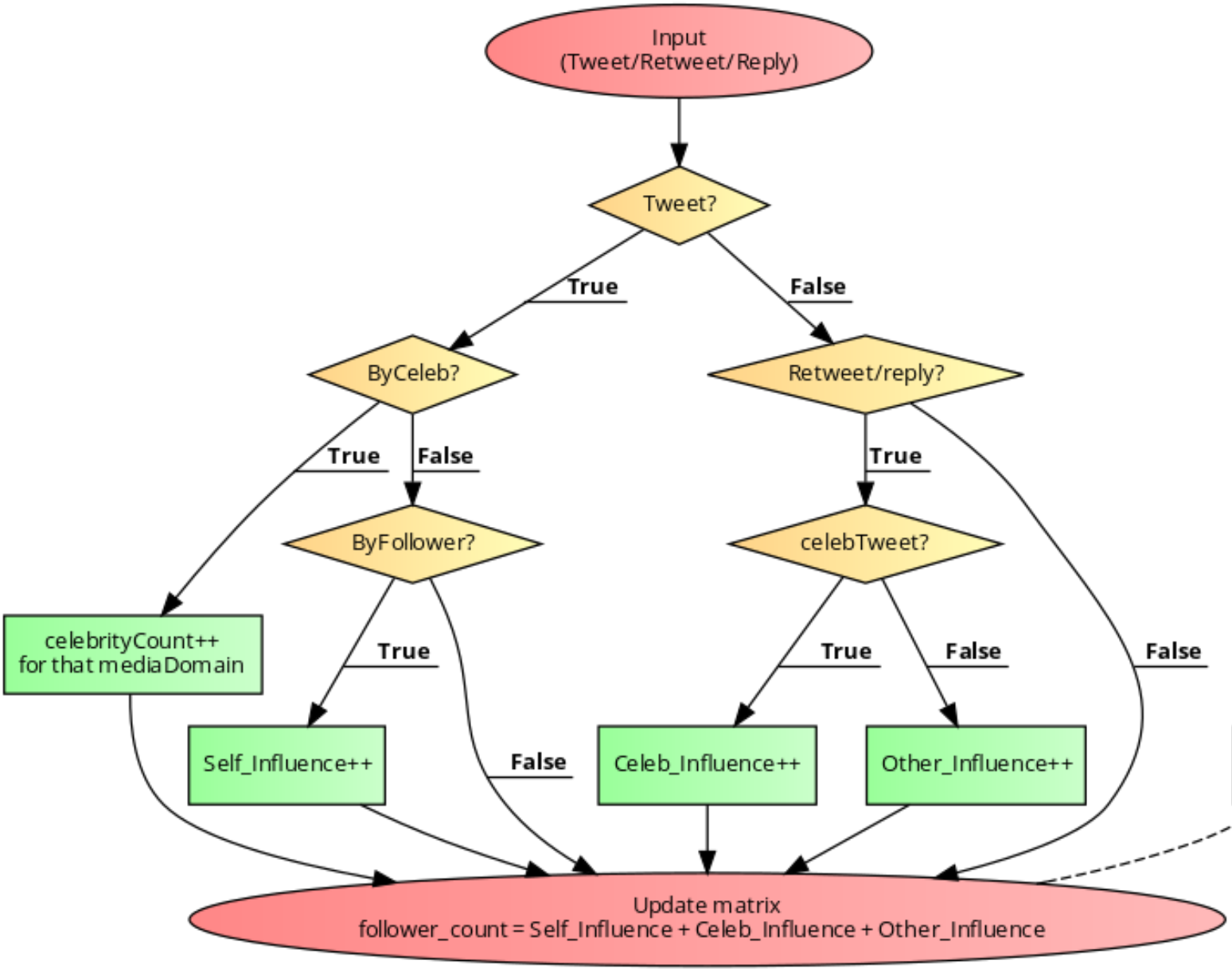
Firstly, we run split.sh on the original files to split them into smaller ones. On these split files, we execute run.sh which , using the java API UrlExpander.jar, extracts the URL from the tweets, expands the domain name and if it matches with the ones provided in mediaDomains, writes the corresponding tweet to the processed output file. The code was optimised for multithreading in order to bring down the exceptionally large processing times caused by the network bottleneck. It took one week to get the final processed files , one corresponding to each of the input files, after running 2000 threads each simultaneously on 3 GCL machines.



Number of files before splitting	After splitting (~ 500mb each)	Time for one file (quad core, 2000 threads)	Total computation time	#Tweets after processing
25	250	20 min	>100 hours	1.3 million

This brought down the database size from above 120GB down to 750MB

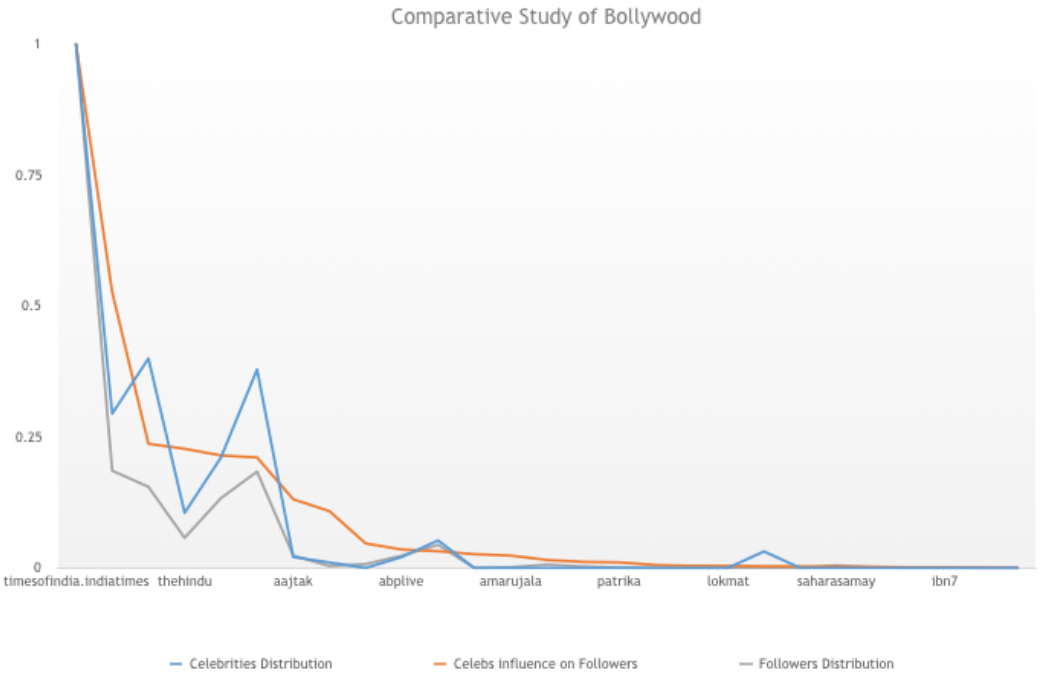
Process FlowChart



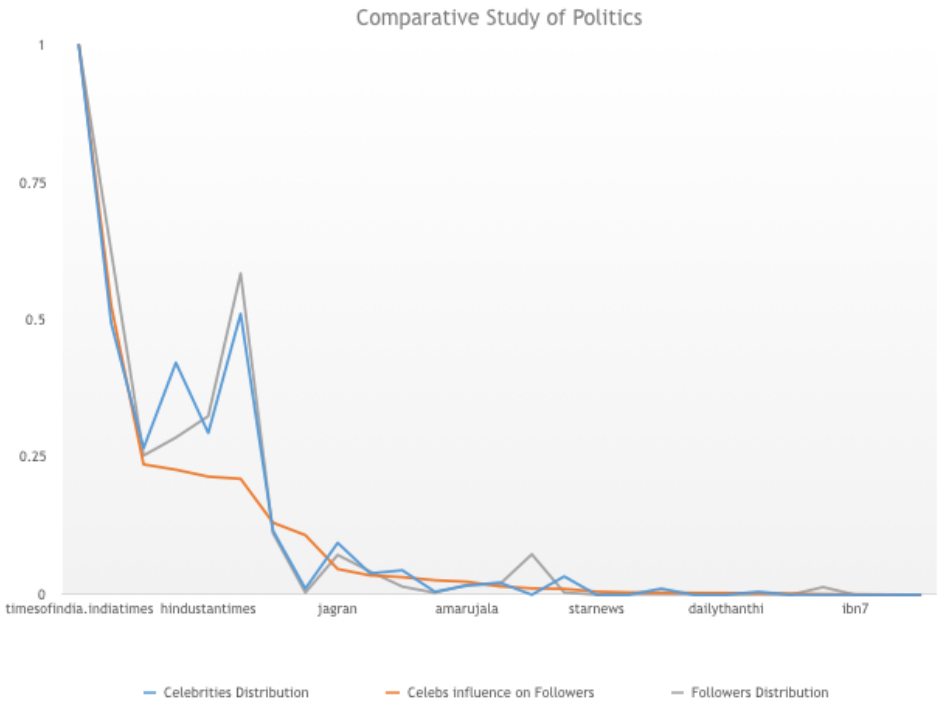
Observations

- **Correlation charts**

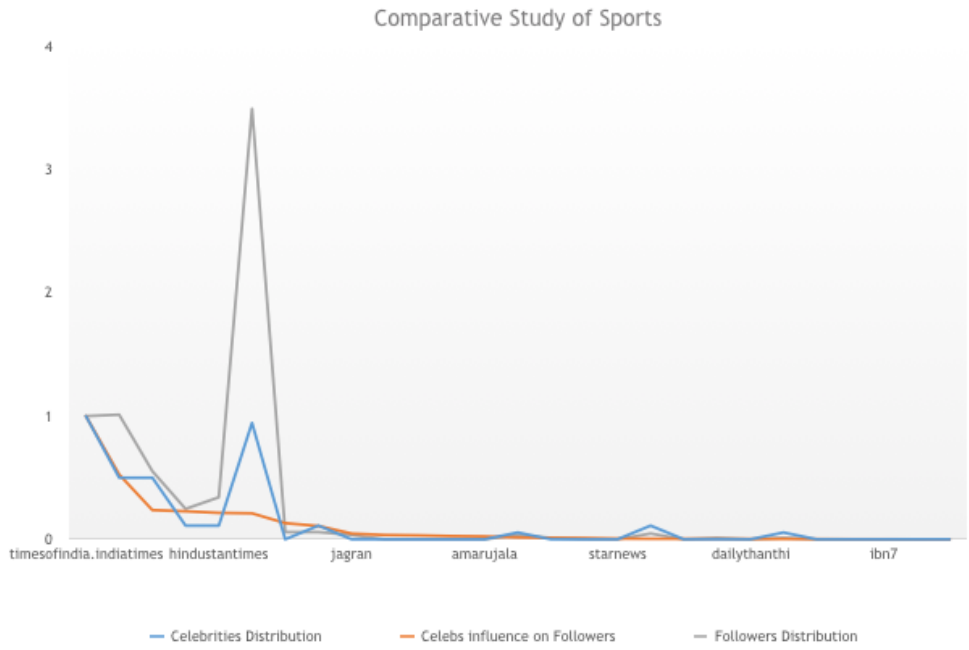
  ~ **Bollywood**



  ~ **Politics**



~ Sports

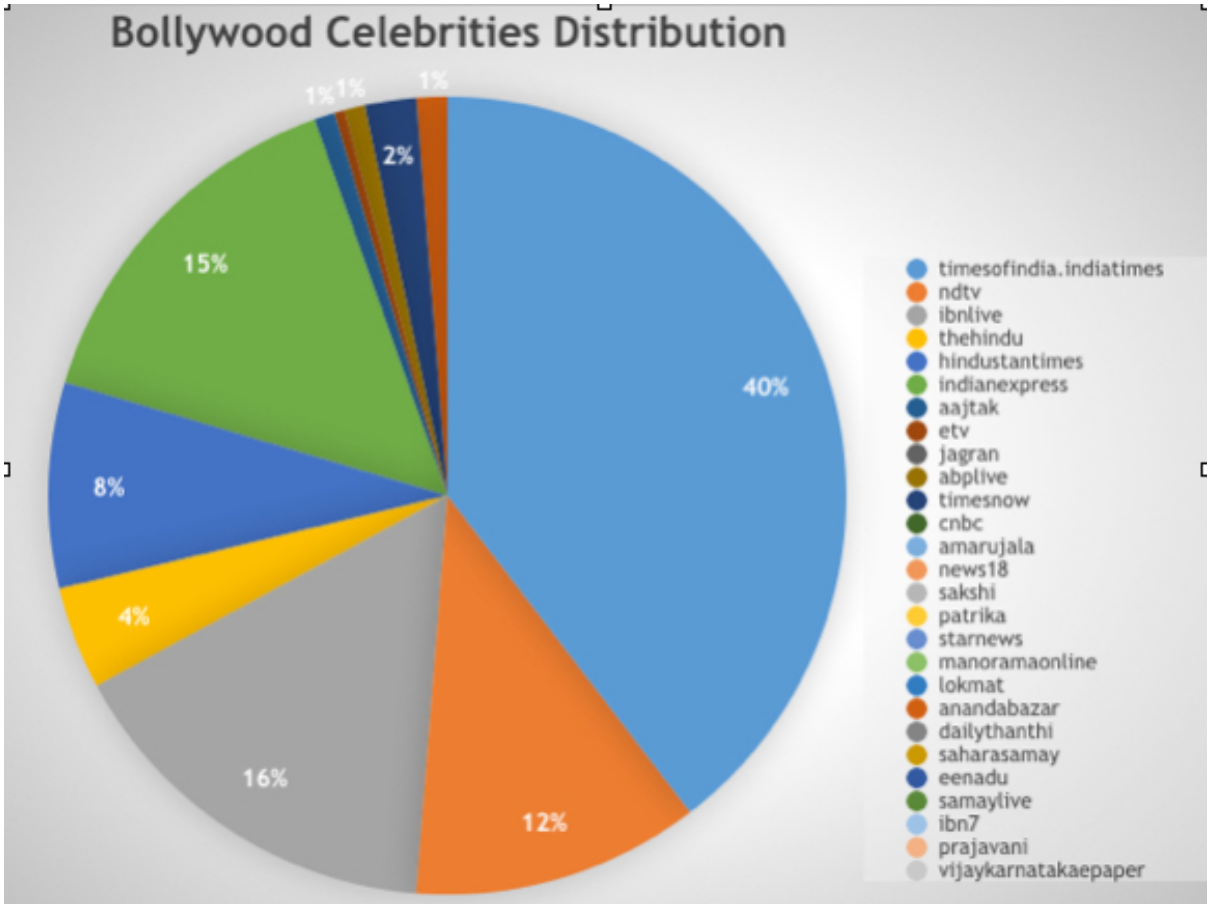


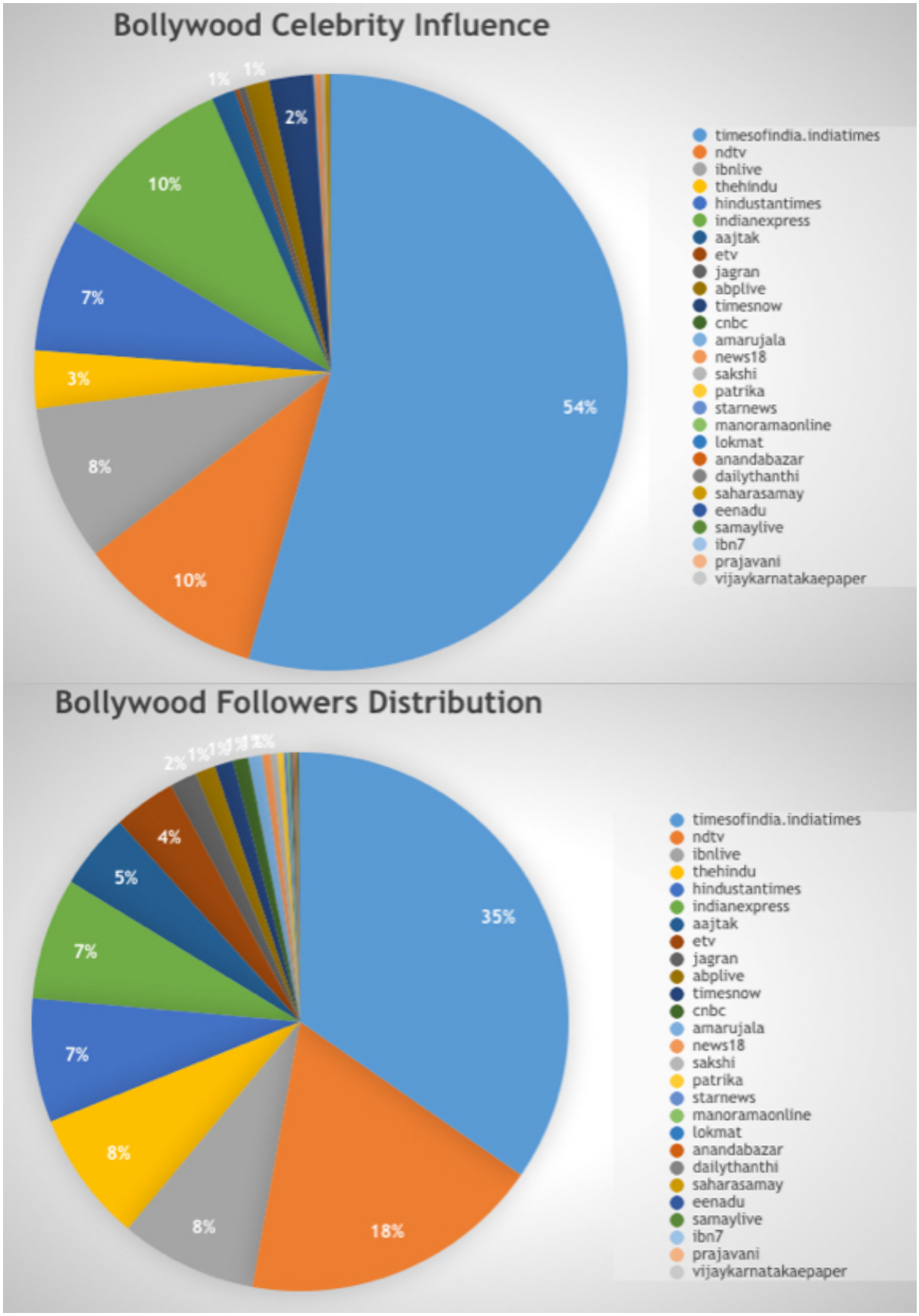
- **Pearson Correlation Test**

Category	Pearson Coefficient (PPMCC)
Bollywood	0.91767161
Politics	0.937015689
Sports	0.673491289

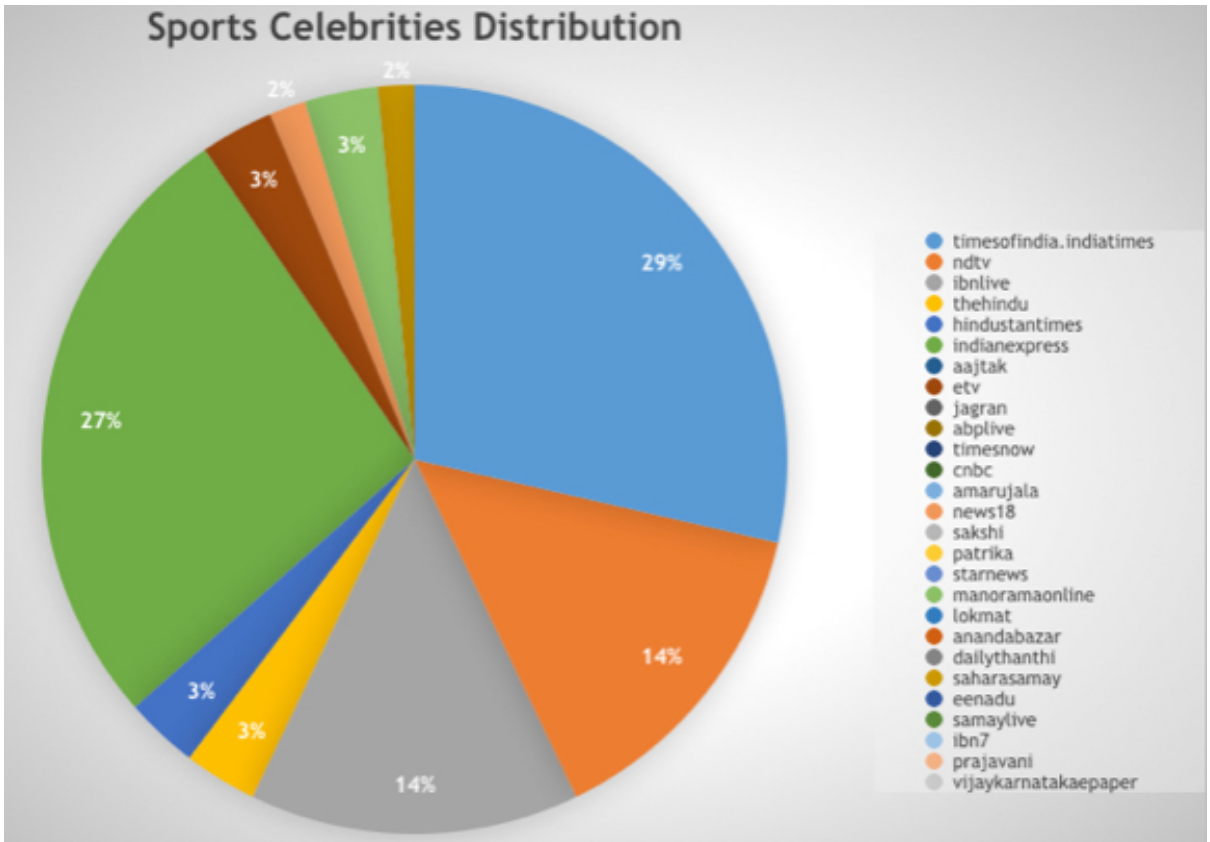
Pie Charts for frequency distribution within each class

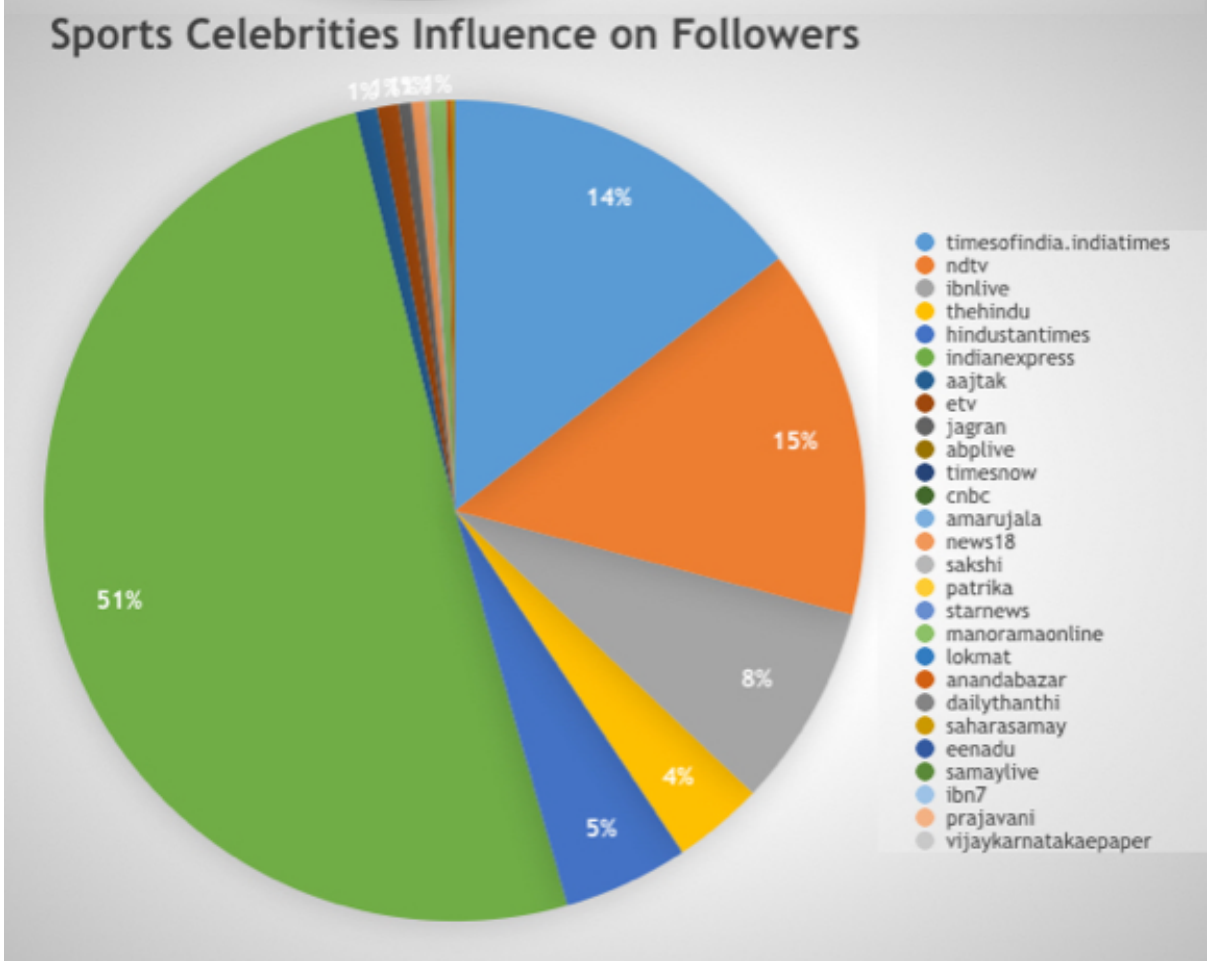
- **Bollywood**

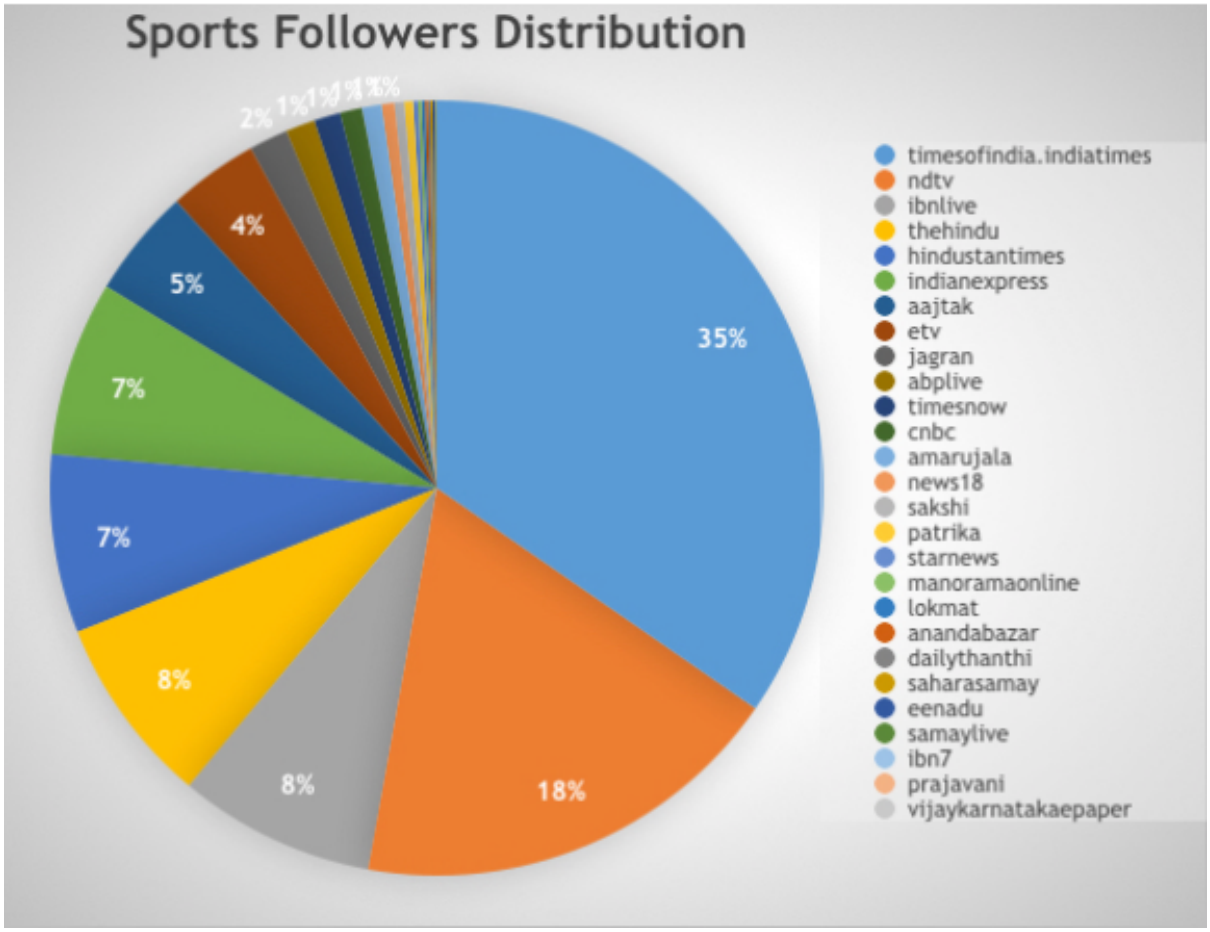

Bollywood Celebrities Distribution
40%
12%
16%
4%
8%
15%
2%
1%
1%
1%
timesofindia.indiatimes
ndtv
ibnlive
thehindu
hindustantimes
indianexpress
aajtak
etv
jagran
abplive
timesnow
cnbc
amarujala
news18
sakshi
patrika
starnews
manoramaonline
lokmat
anandabazar
dailythanthi
saharasamay
eenadu
samaylive
ibn7
prajavani
vijaykarnatakaepaper


Bollywood Celebrity Influence
54%
10%
8%
3%
7%
10%
1%
1%
2%
timesofindia.indiatimes
ndtv
ibnlive
thehindu
hindustantimes
indianexpress
aajtak
etv
jagran
abplive
timesnow
cnbc
amarujala
news18
sakshi
patrika
starnews
manoramaonline
lokmat
anandabazar
dailythanthi
saharasamay
eenadu
samaylive
ibn7
prajavani
vijaykarnatakaepaper
Bollywood Followers Distribution
35%
18%
8%
8%
7%
7%
5%
4%
2%
timesofindia.indiatimes
ndtv
ibnlive
thehindu
hindustantimes
indianexpress
aajtak
etv
jagran
abplive
timesnow
cnbc
amarujala
news18
sakshi
patrika
starnews
manoramaonline
lokmat
anandabazar
dailythanthi
saharasamay
eenadu
samaylive
ibn7
prajavani
vijaykarnatakaepaper

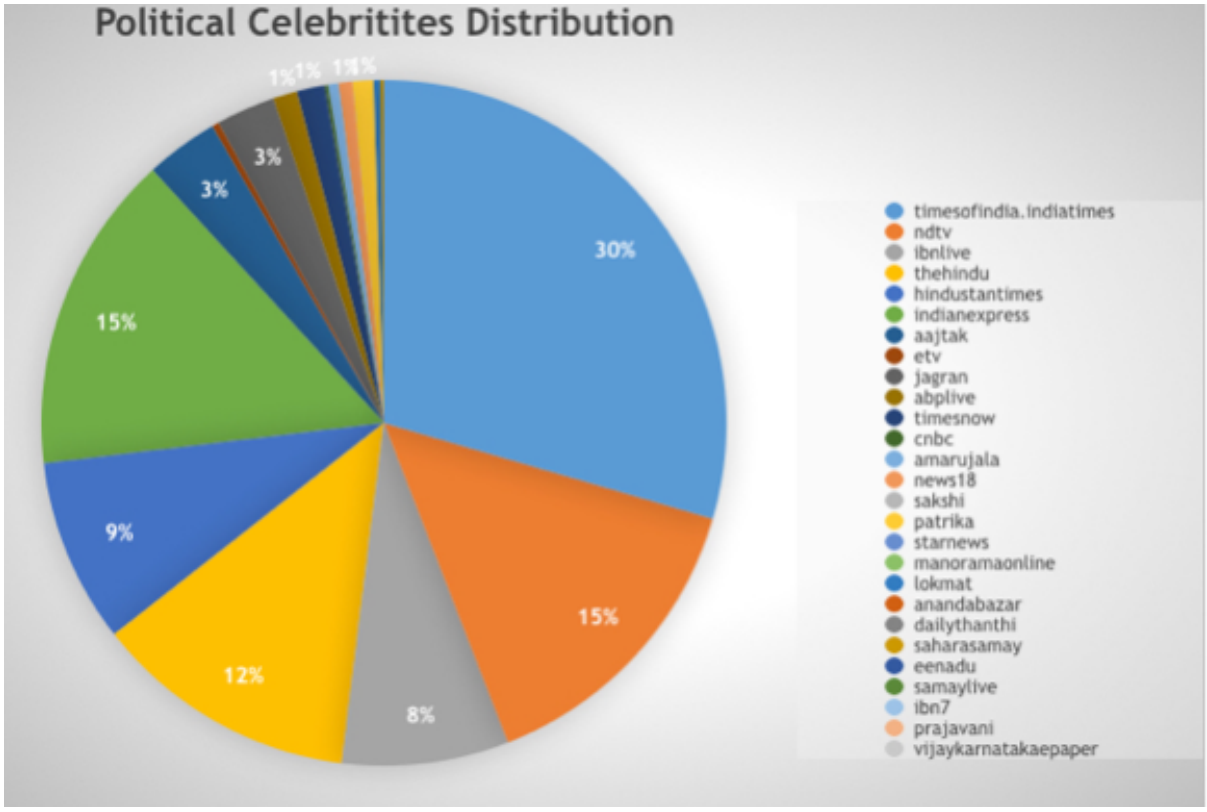
- **Sports**

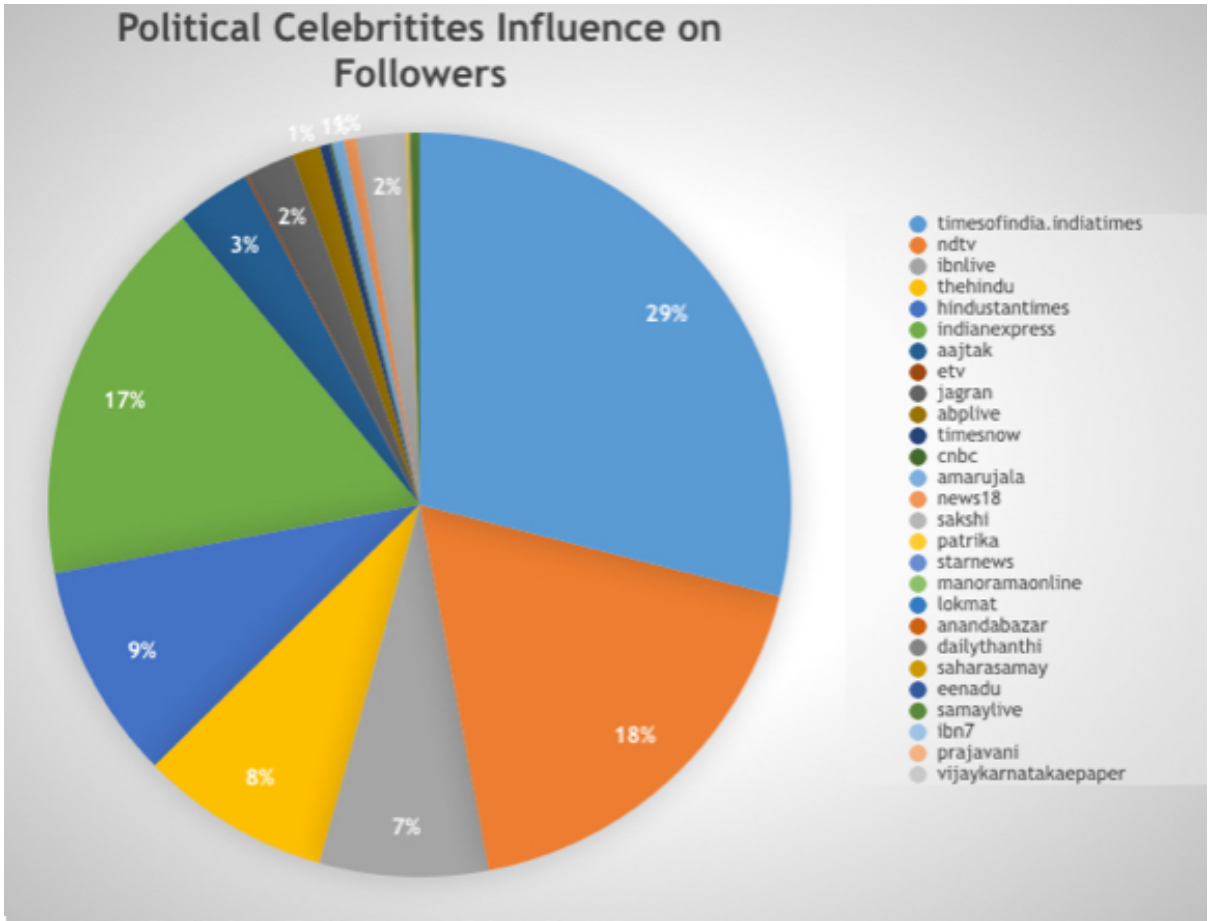

Sports Celebrities Distribution
29%
14%
14%
3%
3%
27%
3%
2%
2%
timesofindia.indiatimes
ndtv
ibnlive
thehindu
hindustantimes
indianexpress
aajtak
etv
jagran
abplive
timesnow
cnbc
amarujala
news18
sakshi
patrika
starnews
manoramaonline
lokmat
anandabazar
dailythanthi
saharasamay
eenadu
samaylive
ibn7
prajavani
vijaykarnatakaepaper

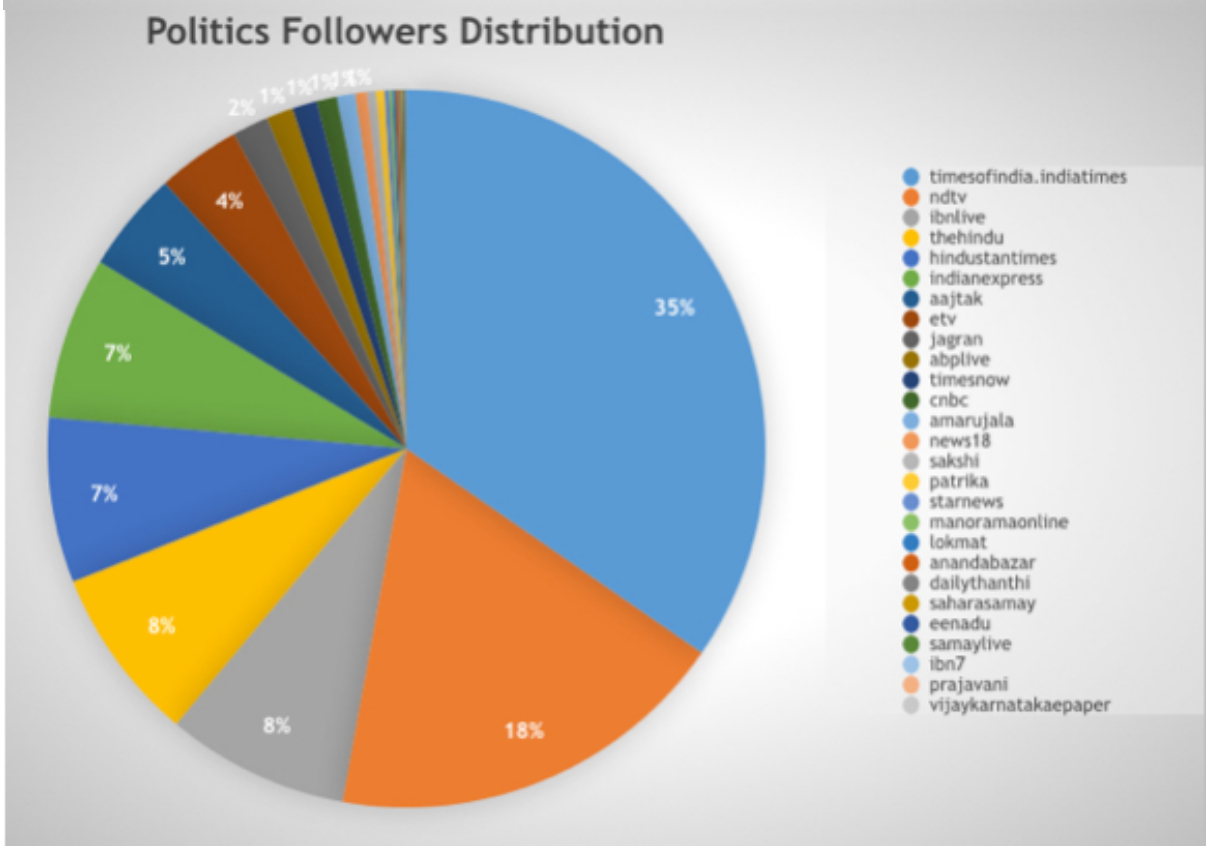

Sports Celebrities Influence on Followers
14%
15%
8%
4%
5%
51%
timesofindia.indiatimes
ndtv
ibnlive
thehindu
hindustantimes
indianexpress
aajtak
etv
jagran
abplive
timesnow
cnbc
amarujala
news18
sakshi
patrika
starnews
manoramaonline
lokmat
anandabazar
dailythanthi
saharasamay
eenadu
samaylive
ibn7
prajavani
vijaykarnatakaepaper


Sports Followers Distribution
35%
18%
8%
8%
7%
7%
5%
4%
2%
timesofindia.indiatimes
ndtv
ibnlive
thehindu
hindustantimes
indianexpress
aajtak
etv
jagran
abplive
timesnow
cnbc
amarujala
news18
sakshi
patrika
starnews
manoramaonline
lokmat
anandabazar
dailythanthi
saharasamay
eenadu
samaylive
ibn7
prajavani
vijaykarnatakaepaper

- **Politics**


Political Celebritites Distribution
30%
15%
12%
9%
15%
3%
3%
timesofindia.indiatimes
ndtv
ibnlive
thehindu
hindustantimes
indianexpress
aajtak
etv
jagran
abplive
timesnow
cnbc
amarujala
news18
sakshi
patrika
starnews
manoramaonline
lokmat
anandabazar
dailythanthi
saharasamay
eenadu
samaylive
ibn7
prajavani
vijaykarnatakaepaper


Political Celebritites Influence on Followers
29%
18%
7%
8%
9%
17%
3%
2%
2%
timesofindia.indiatimes
ndtv
ibnlive
thehindu
hindustantimes
indianexpress
aajtak
etv
jagran
abplive
timesnow
cnbc
amarujala
news18
sakshi
patrika
starnews
manoramaonline
lokmat
anandabazar
dailythanthi
saharasamay
eenadu
samaylive
ibn7
prajavani
vijaykarnatakaepaper


Politics Followers Distribution
35%
18%
8%
8%
7%
7%
5%
4%
timesofindia.indiatimes
ndtv
ibnlive
thehindu
hindustantimes
indianexpress
aajtak
etv
jagran
abplive
timesnow
cnbc
amarujala
news18
sakshi
patrika
starnews
manoramaonline
lokmat
anandabazar
dailythanthi
saharasamay
eenadu
samaylive
ibn7
prajavani
vijaykarnatakaepaper

Limitations

- The number of tweets finally extracted from the original database should be greater. Currently out of 450m tweets, 1.3m are extracted into the processed database
- Additional constraints and factors like overlap of followers across categories need to be carefully handled
- A simple cause and effect relationship shouldn't be assumed since celebrities themselves may be followers and be influenced by others

Inferences

- Political celebrities are more active in sharing media sources and so are their followers
- Bollywood and Sports celebrities behave more like the general public in their media choices, as hypothesised
- Therefore, both the pie charts and the line charts indicate affinity of the political followers' choices to that of their celebrities
- Reading habits of the followers do resemble those of the celebrities

Conclusion

On the basis of our observations, we conclude that celebrities , except for political ones, behave more like the general public, though they do influence their followers albeit to a much lower extent that Politicians. This holds true to our hypotheses since apart from being the only zone for subjectivity in the news reports, political celebrities have a personal interest vested in being more active with respect to news and media houses and affinity for those news which tend to favour their political agendas. This in turn skews the choices of their followers for the same reasons.