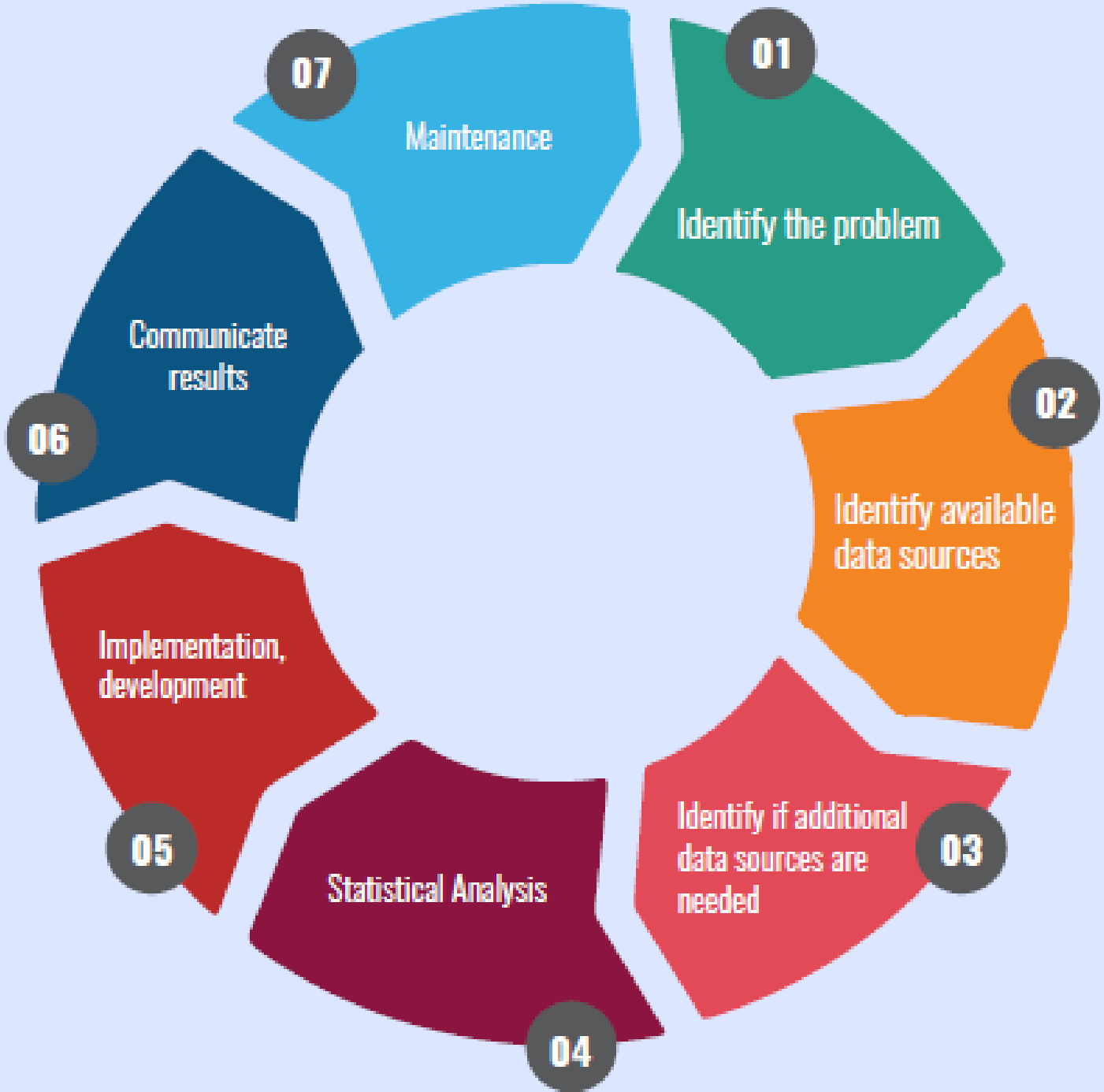
Golden Data Overview

WHY

- Gather rich data using features which impact the target (Views)
- Model can learn better and gives better results.
- More Accurate Inference



- While creating the golden dataset our aim was to describe the video as much as possible.
- For that, we selected features related to Text, Visual/Image, Audio, Video Descriptive and Date-Time.
- We have created the golden sample dataset of 1449 videos
- All the videos are taken from NEWJ main page (Facebook)

DATA

Text

- **Title**
- **Social Copy**
- **Tags**

Visual/Image

- **visual_first_3_seconds**
- **thumbnail**
- **thumbnail_Sentiment**

Audio

- **background_music_type_first_3_seconds**
- **voice_first_3_seconds**

DATA

Video Descriptive

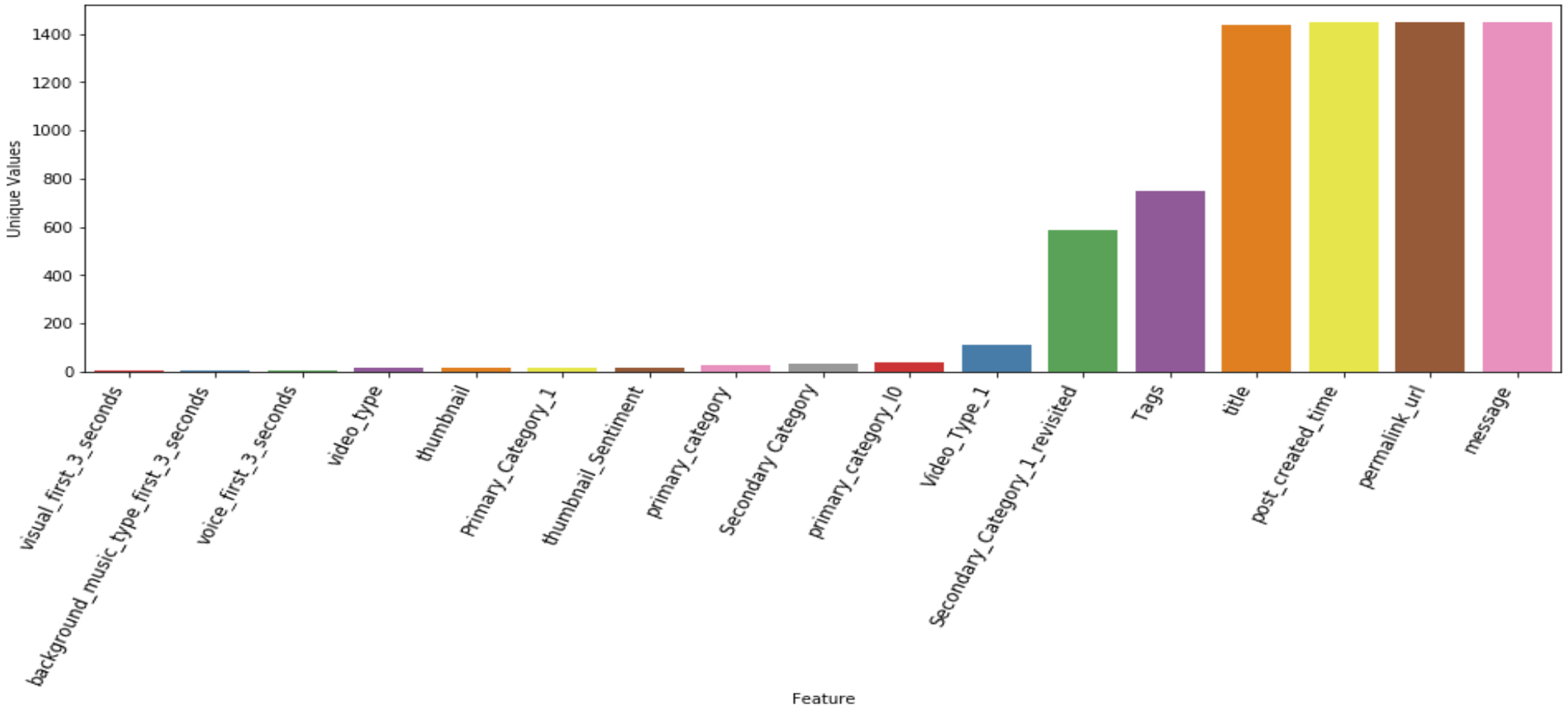
- **Primary Category**
- **Secondary Category**
- **Video Type**

Other

- **Date**
- **Time**

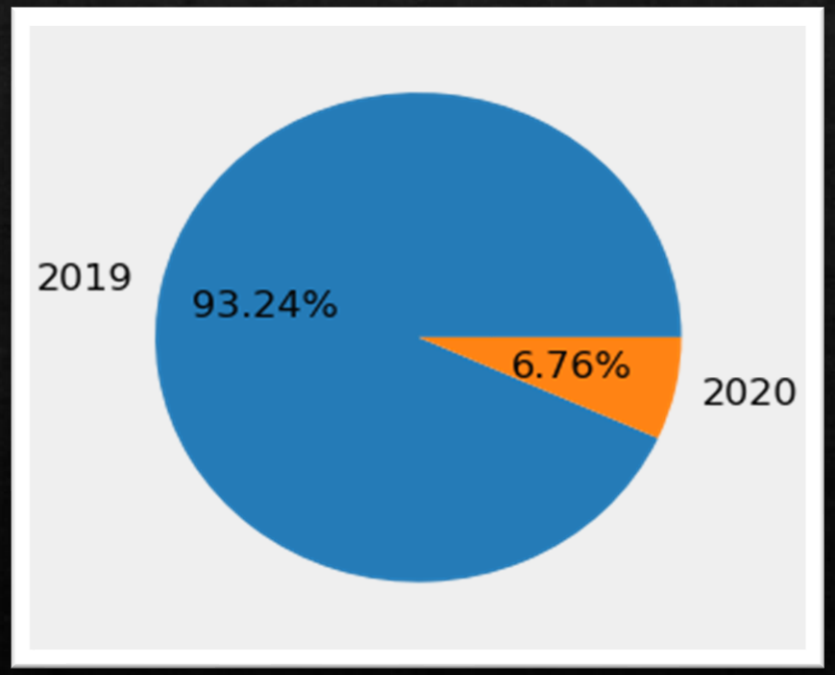
FEATURES VALUE DISTRIBUTION

As expected, features like permalink, message (social copy), post created time, title are unique. Categorical features should have limited unique values (less). Whereas in our dataset most of them have limited unique values except Tags, Secondary_Category_1_revisited, Video_Type_1. Though, it is obvious to have more unique tags. But we need to investigate the Secondary_Category_1_revisited and Video_Type_1.



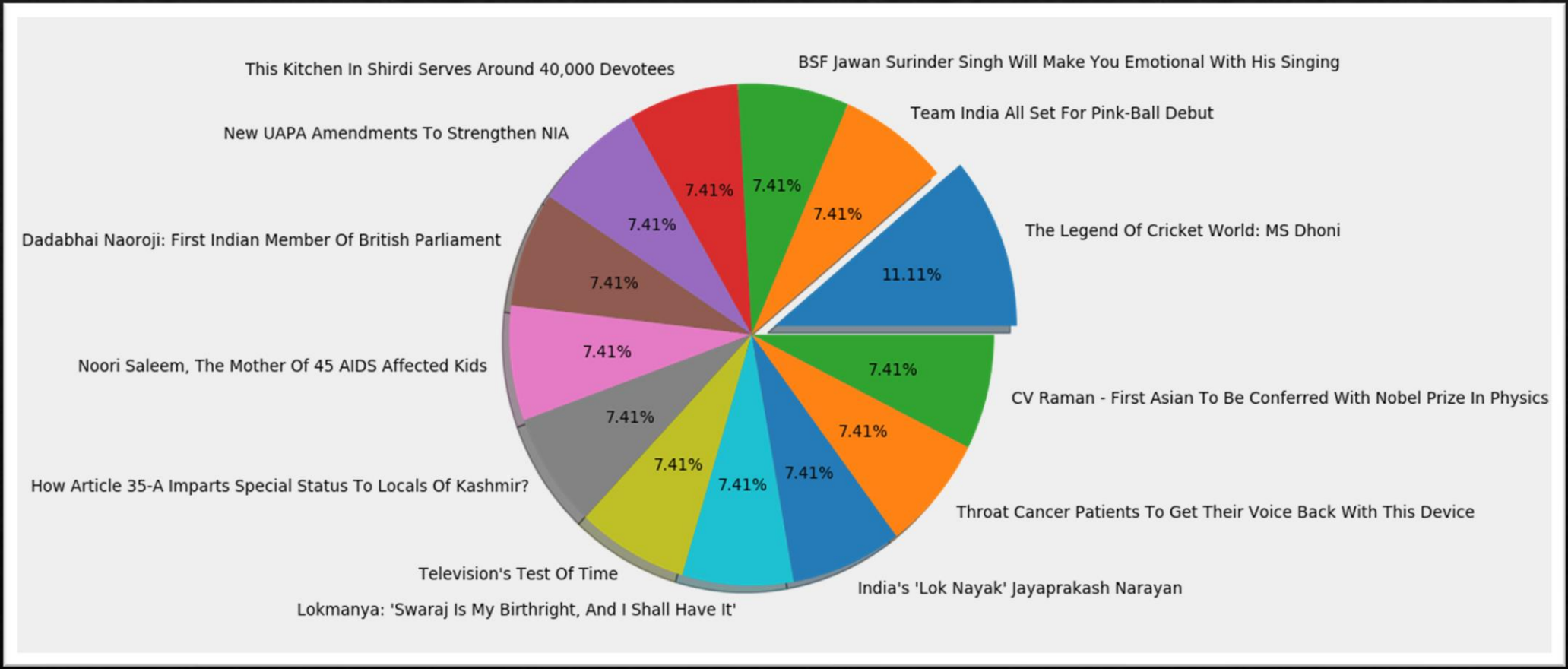
Date/Time

- Our data contain videos from mid of May-2019 till mid of April-2020
- Almost 7% of videos are from 2020 and remaining from 2019



Title & Social Copy

- **All the social copy are unique.**
- **13 titles are duplicated**
- **All of them used twice except ‘The Legend Of Cricket World: MS Dhoni’ which was repeated for 3 times**



Primary Category

***Definition**:* Primarily overall video falls into which category

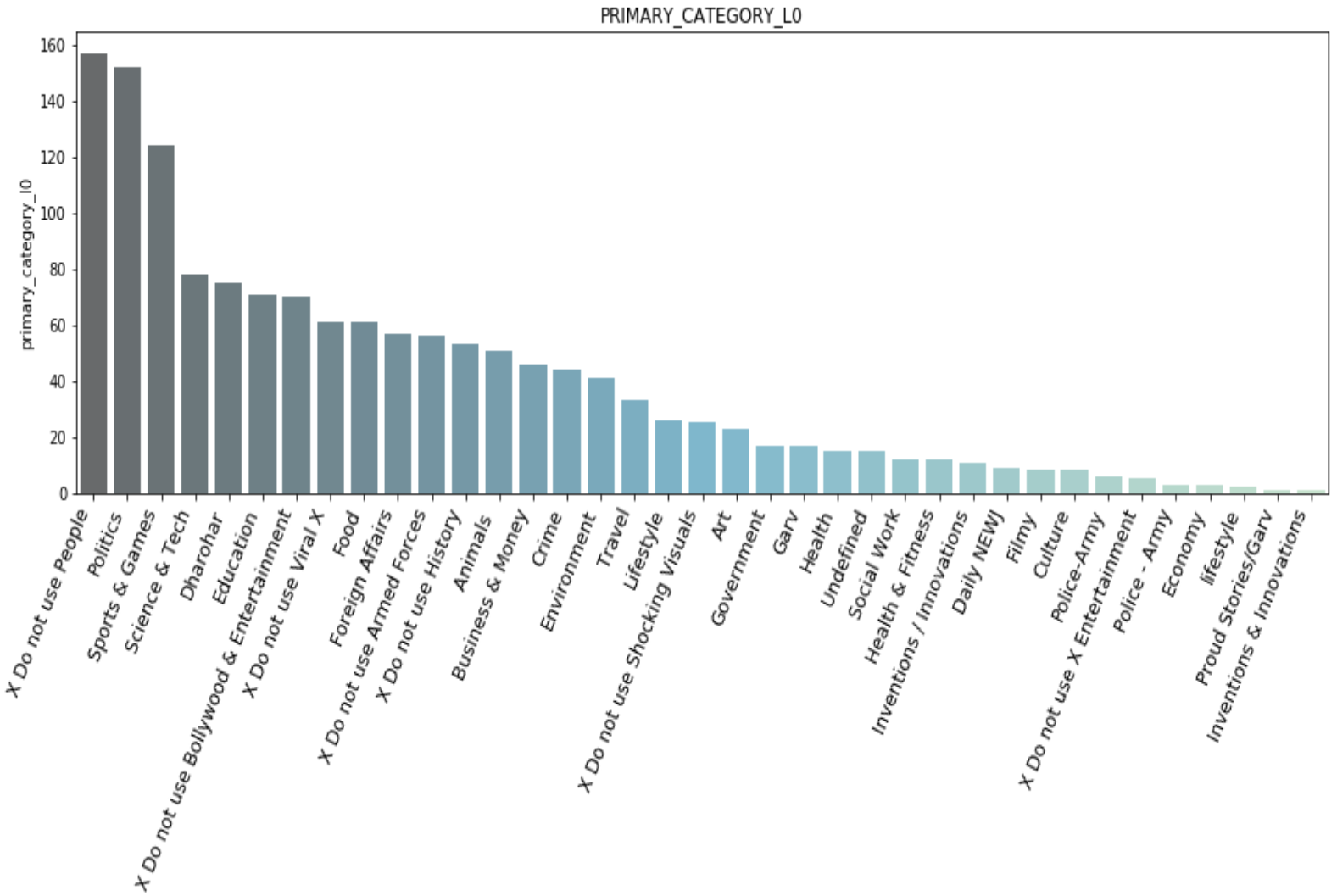
***Example**: Suppose in the video, if Sachin Tendulkar is talking about his food habits and daily routine, then Primary category will be Lifestyle because video is all about Lifestyle.*

primary_category_l0	primary category	primary_category_1
As per old categorization	Correction in old distribution	New categorization
`37 categories are used`	`28 categories are used`	`16 categories are used`

Primary Category

As Per Old Categorization

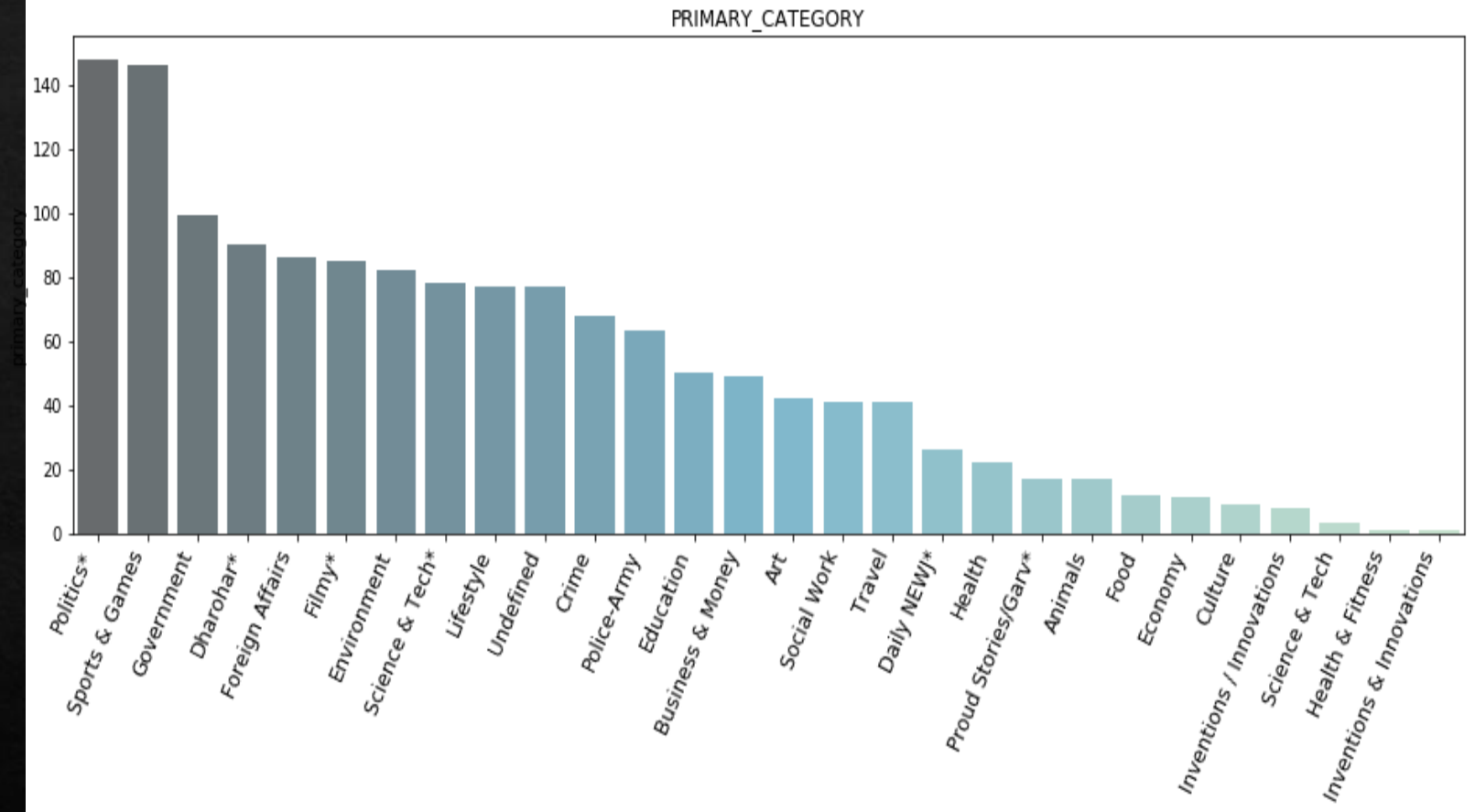
Data is badly distributed along all the features



Primary Category

As Per Corrections In Old Categorization

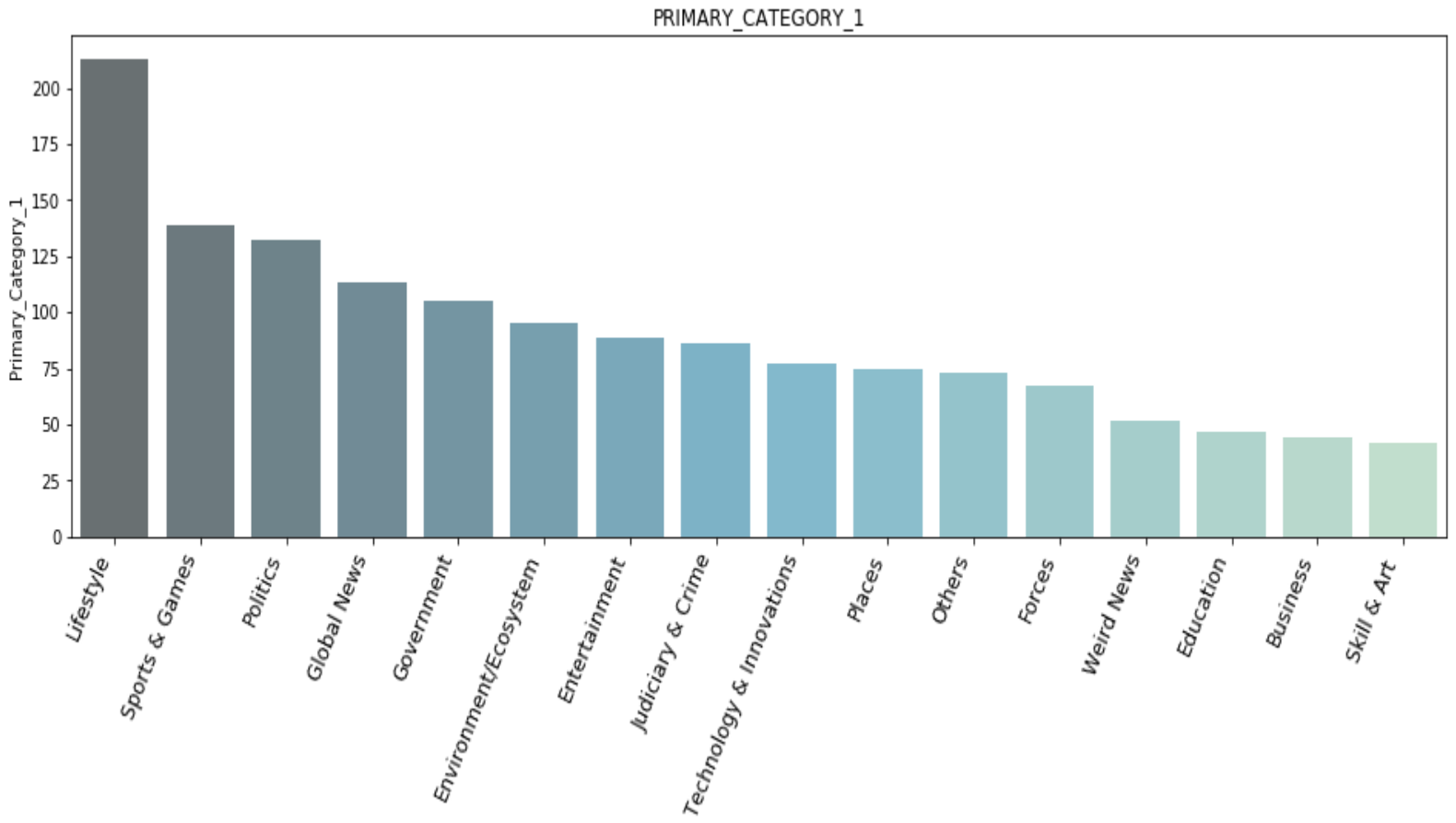
Data distribution got better after correction, still it is skewed



Primary Category

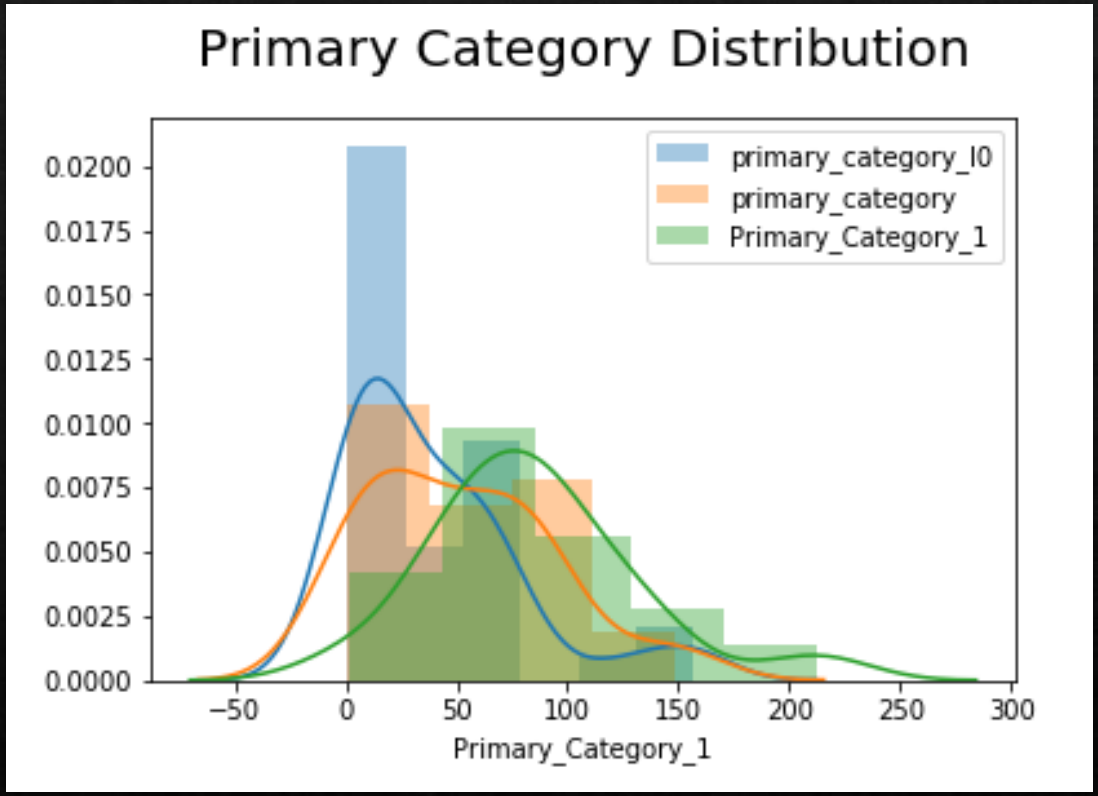
As Per New Categorization

We were able to decrease the number of categories



Primary Category

- “More The Bell Like Curve, Better The Distribution”
- Distribution is better as per new categorization (Primary_Category_1) compare to correction in old categorization
- Where correction in old categorization got better distribution compare to old one.



Secondary Category

Definition: What are the other subcategories video can fall into

Example: For same example from primary category, secondary category will be *cricket* because there is famous person whose profession is cricket, then other subcategories will be *famous personality, food.*

Secondary Category

- As per old categorization
- 33 Categories were used

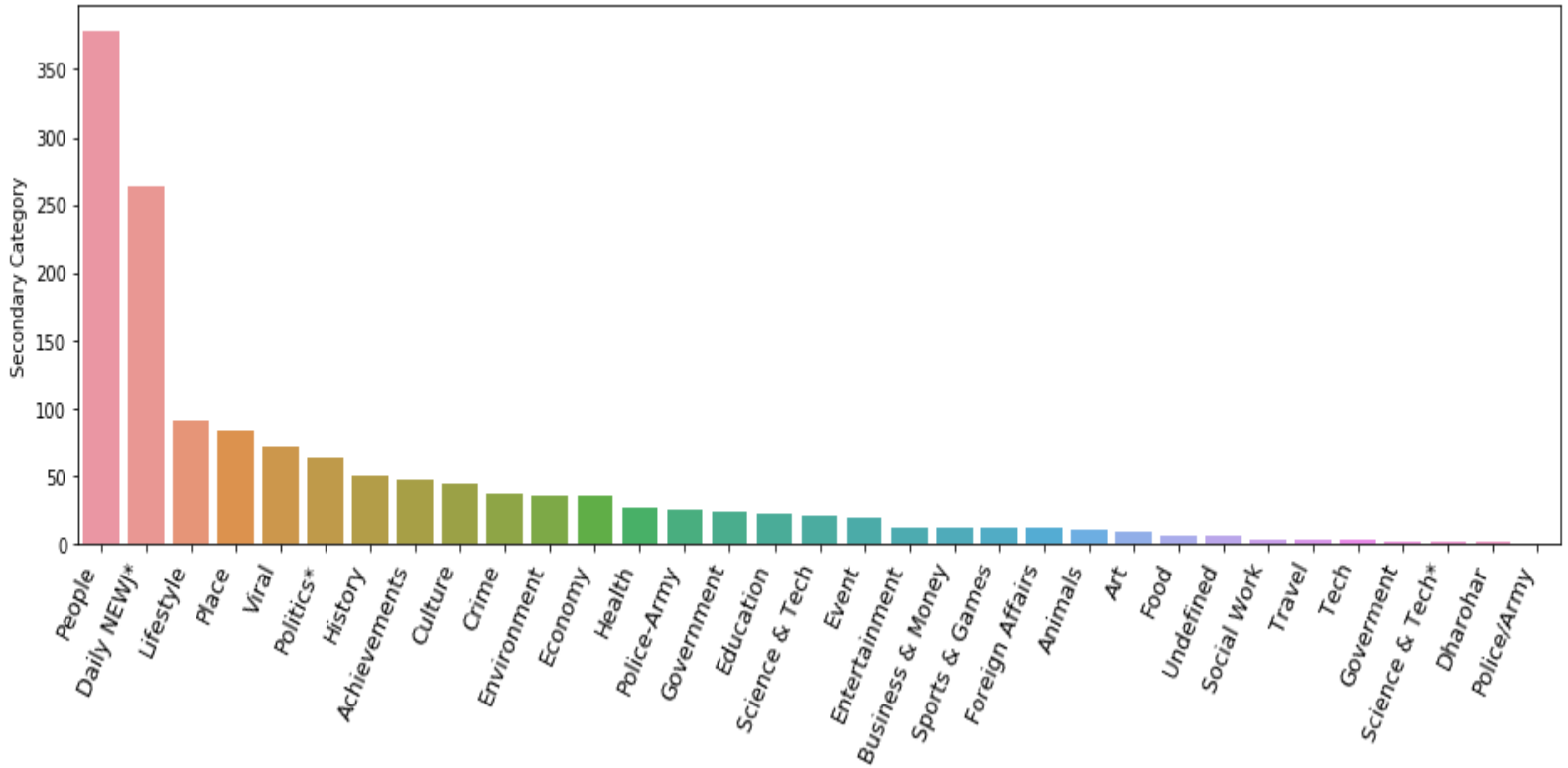
Secondary Category Revisited

- As per new categorization
- In new categorization, we can have up to three secondary categories.
- It is because one video can fall into several subcategories.
- That is why we can see 585 categories, but the actual number is much lesser.

Secondary Category

As Per Old Categorization

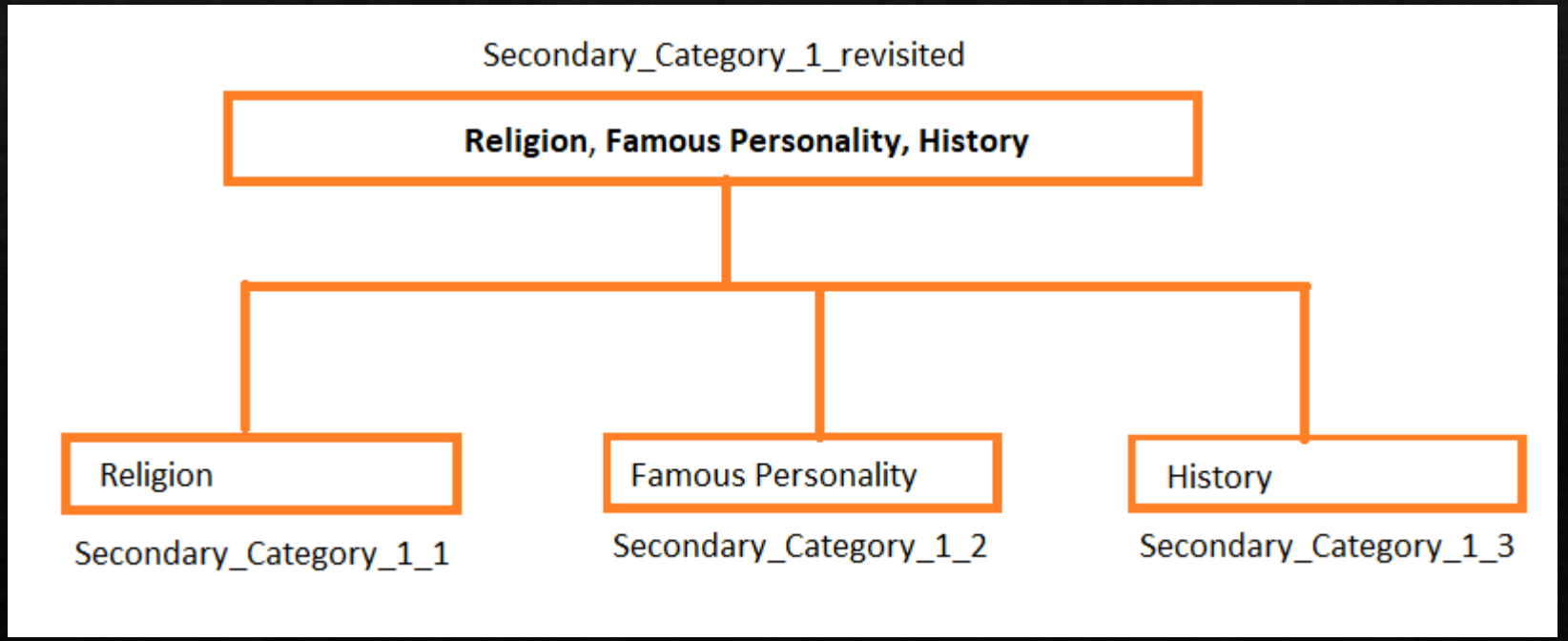
Badly distributed along all the categories



Secondary Category Revisited

We can have up to 3 Secondary categories separated by comma. Here, to see exact number of secondary categories, to check the distribution and for the modelling purpose we have created three different columns for Secondary_Category_1_revisited.

For example:

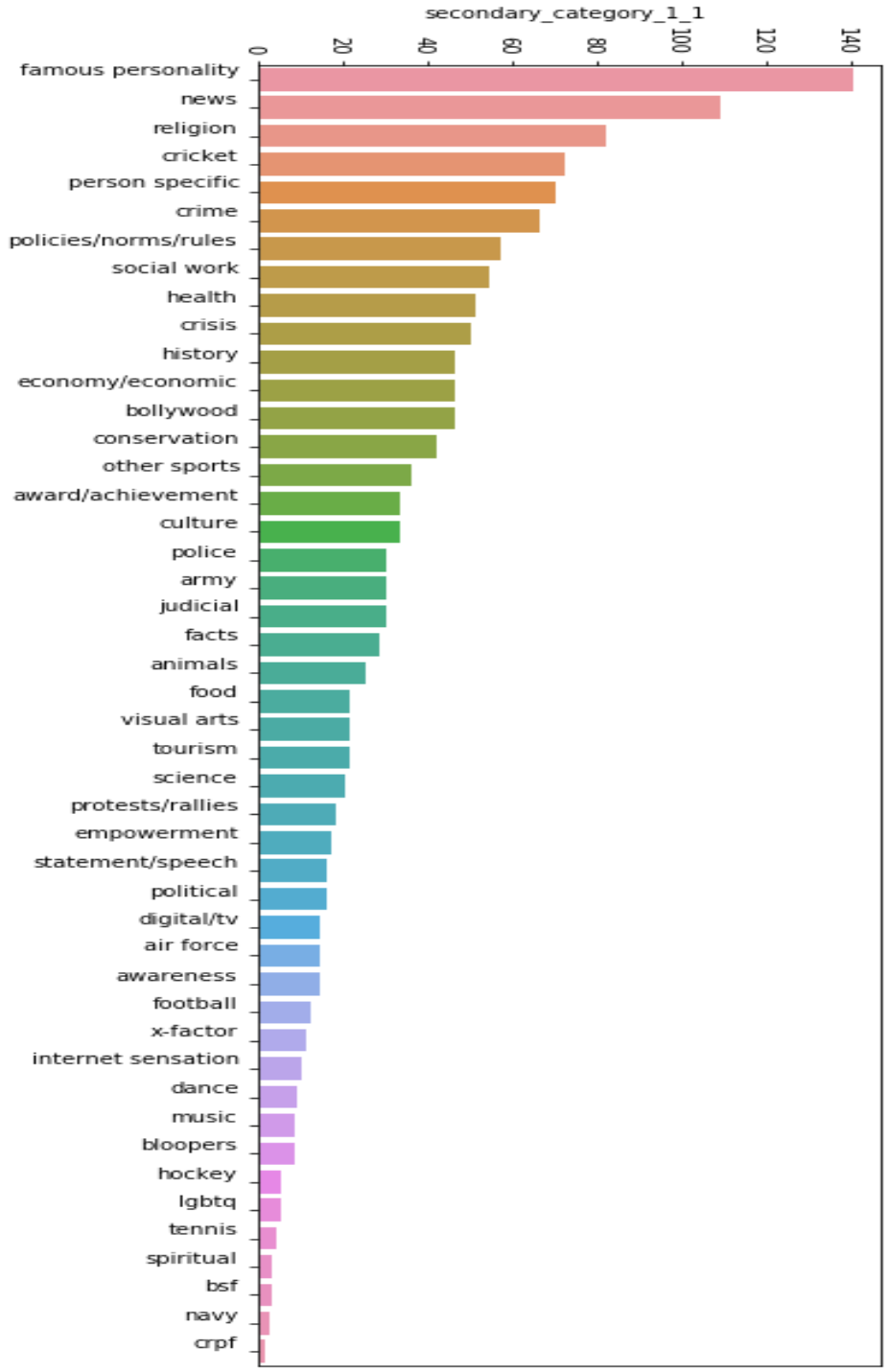


Secondary_Category_1_1

As Per New Categorization

To find actual number of categories and to make it easier and better for machine to interpret and make better predictions, we have created three different columns.

The bar graph on the right side is for the Secondary_Category_1_1.



Secondary_Category_1_2

no_s_category means no secondary category. It means video have one 1 secondary category

Blank space (‘ ’) is because of extra commas.

So, it means no new category is used which is good.

Near about 500 videos do not have two secondary categories

```
Unique Categories From Secondary Category 1:

{'', 'no_s_category'}
```

In [92]:

Secondary_Category_1_3

no_s_category means no secondary category. It means video have one 1 secondary category

Blank space (‘ ’) is because of extra commas.

So, it means no new category is used which is good.

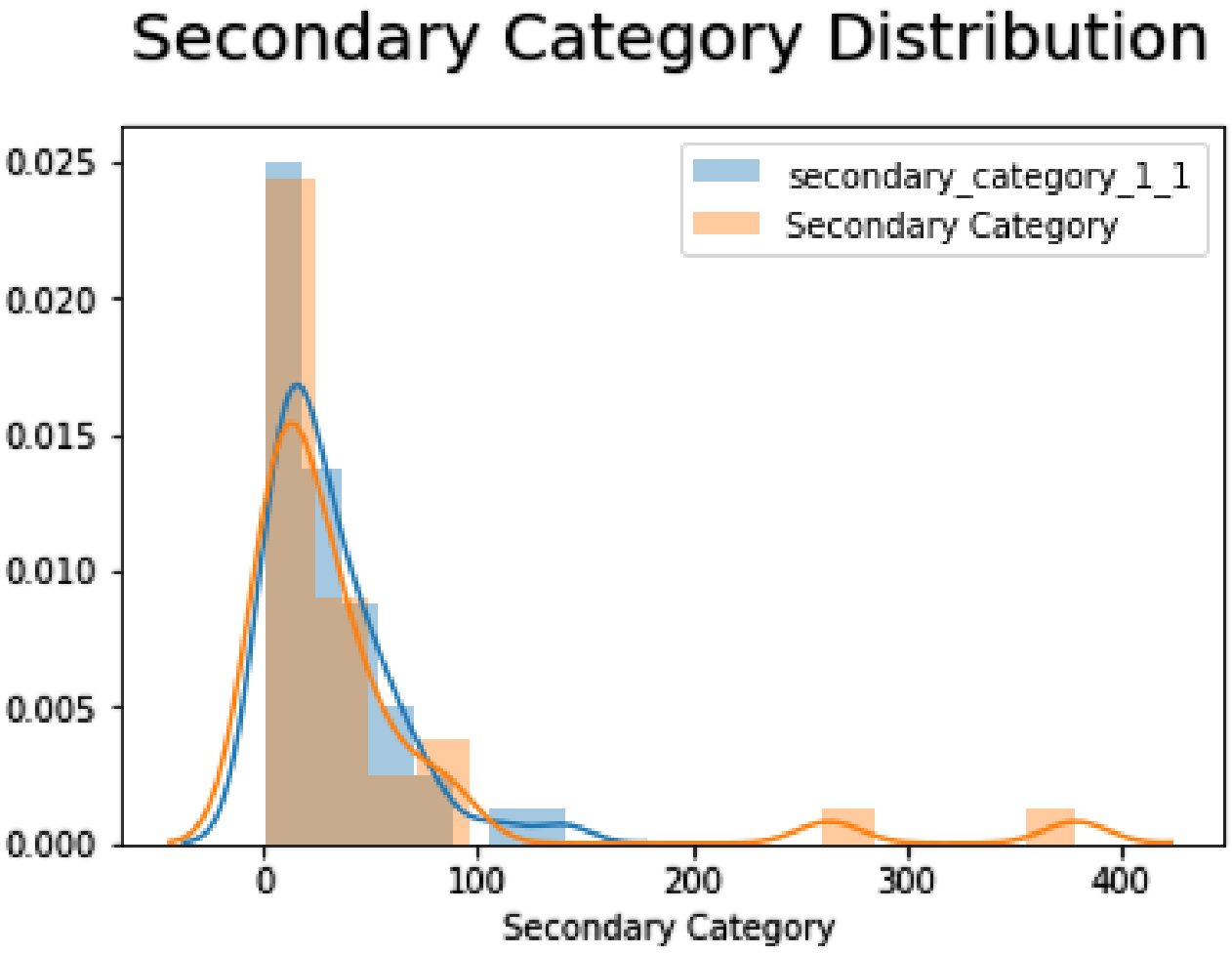
Near about 1200 videos do not have three secondary categories

```
Unique Categories From Secondary Category 1:

{'', 'no_s_category'}
```

Comparison Between Old Secondary Category & New Secondary Category

Distribution got better in the new categorization but not by much.



Video Type

Video Type

- As per old categorization
- 33 Categories were used

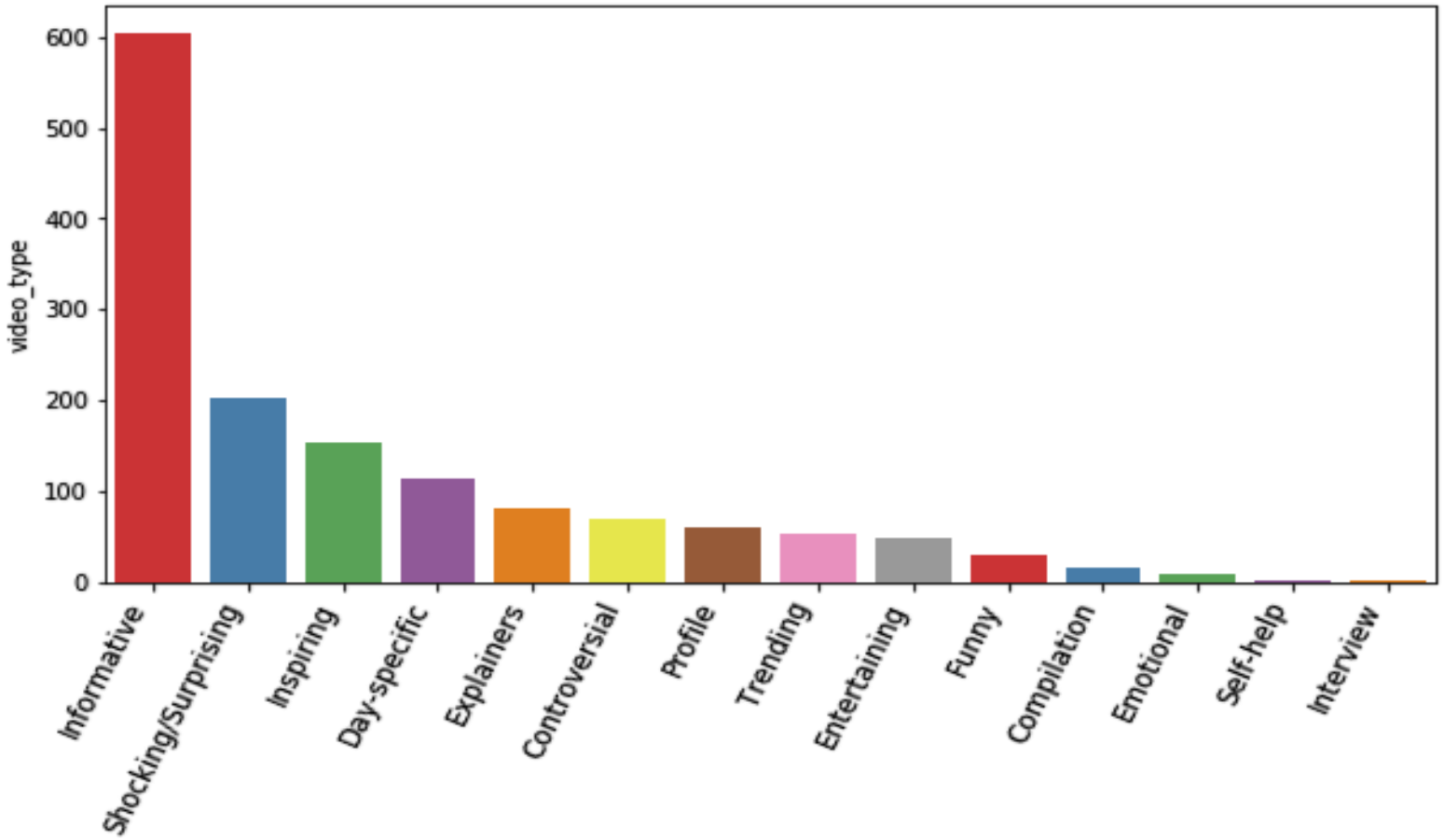
Video Type 1

- As per new categorization
- In new categorization, we can have up to two video types
- It is because one video can fall into several types.
- This is because we can see 107 categories, but the actual number is much lesser

Video Type

As Per Old Categorization

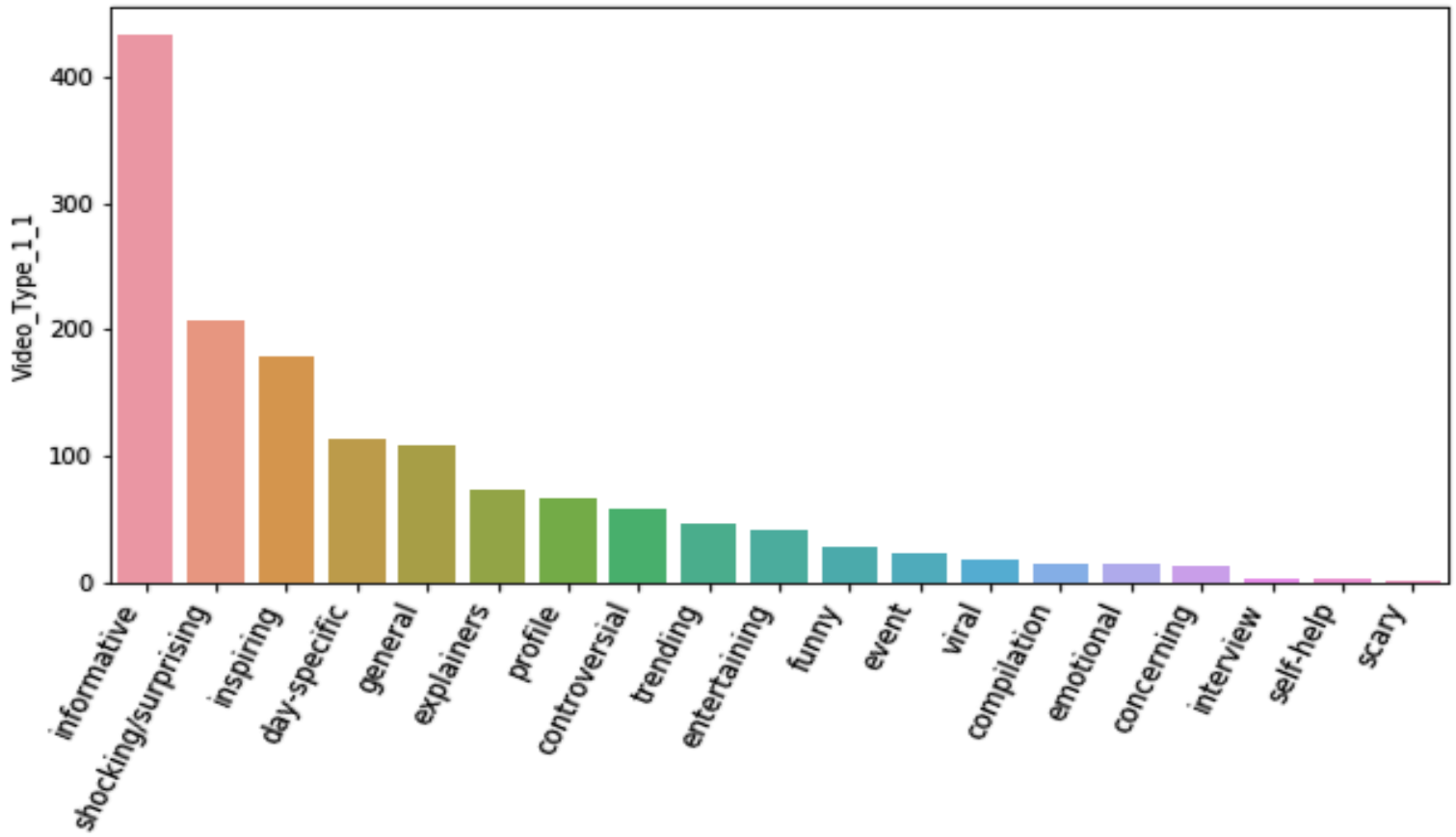
Distribution of data is skewed



Video Type

As Per New Categorization

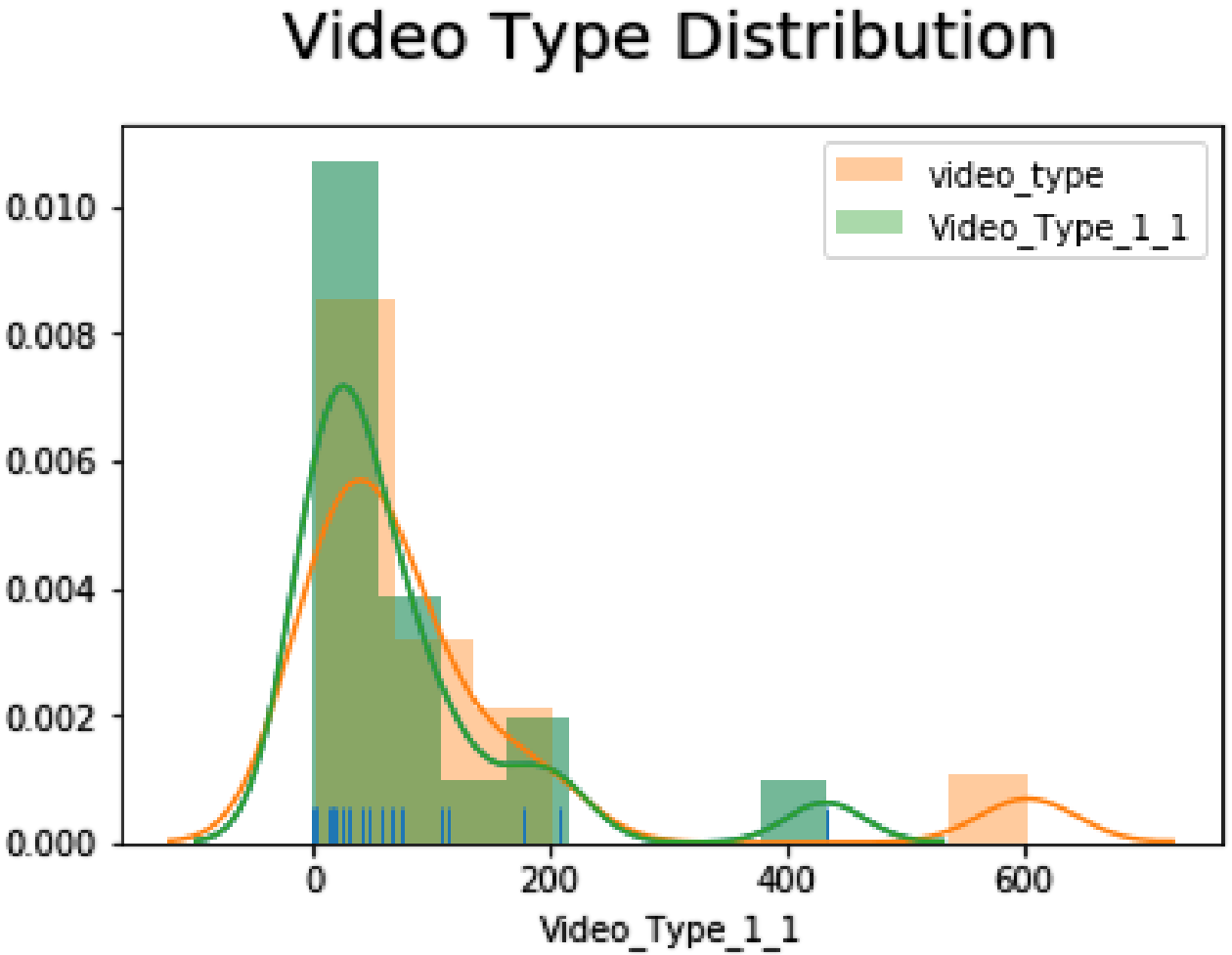
Still the distribution is skewed



Video Type Conclusion

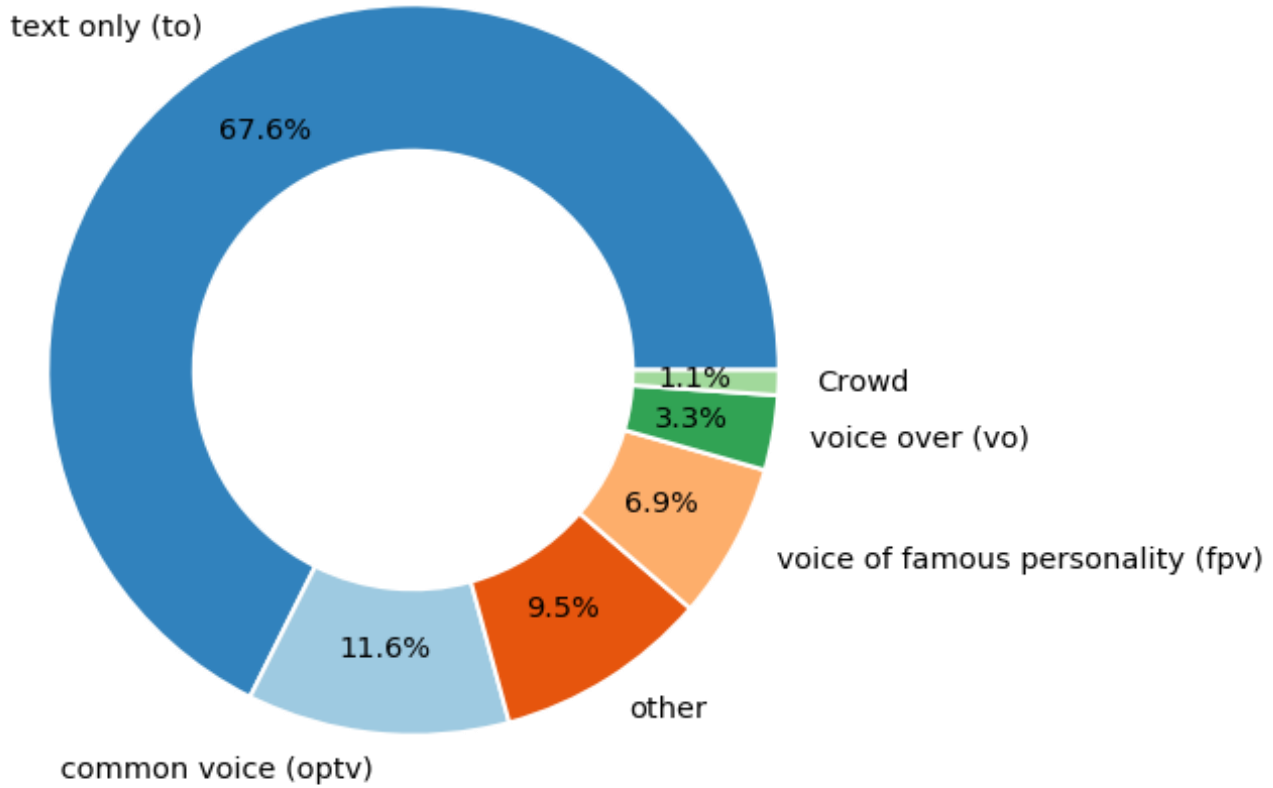
As we can see, in new distribution we were not able to change the distribution by much.

We need to check, if there are some mistakes in categorization (what are the changes required)



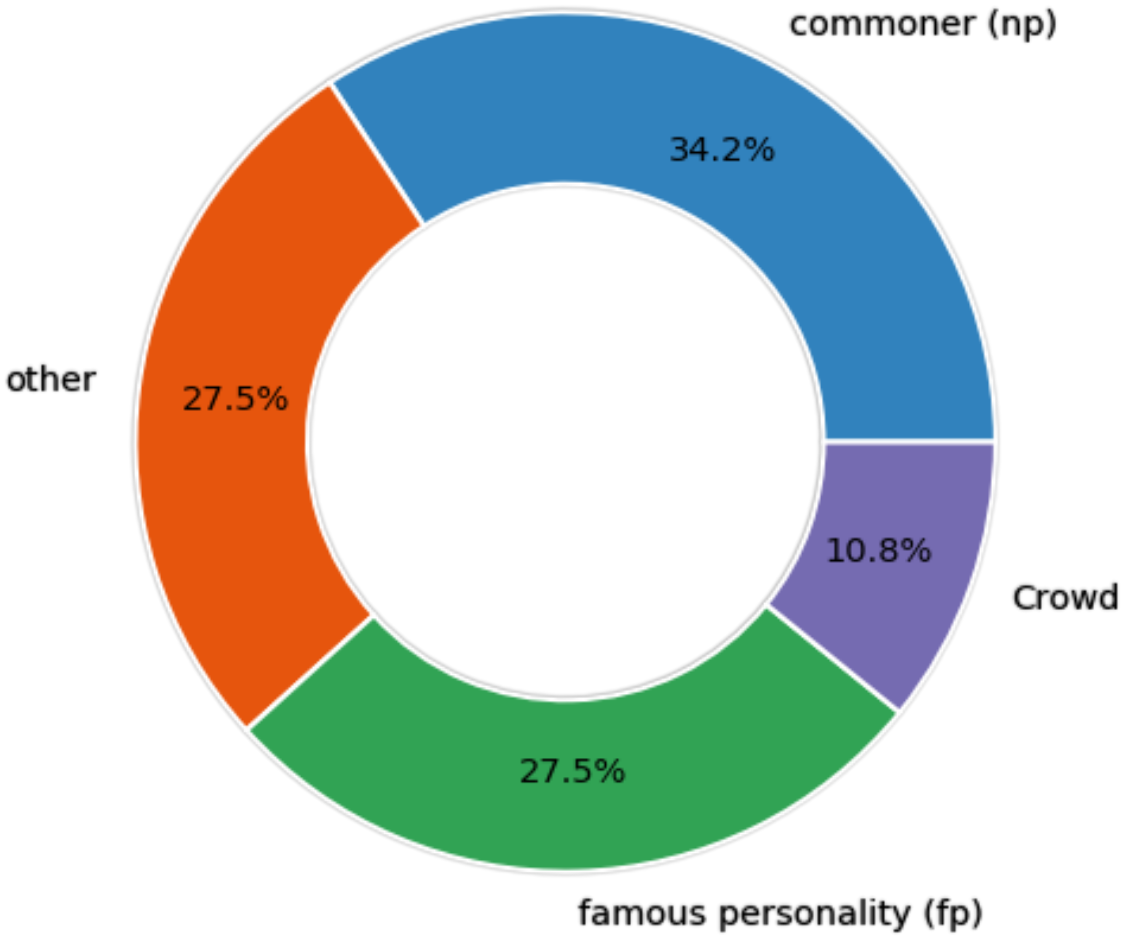
voice_first_3_seconds

- More than 67 percent of videos starts with text only



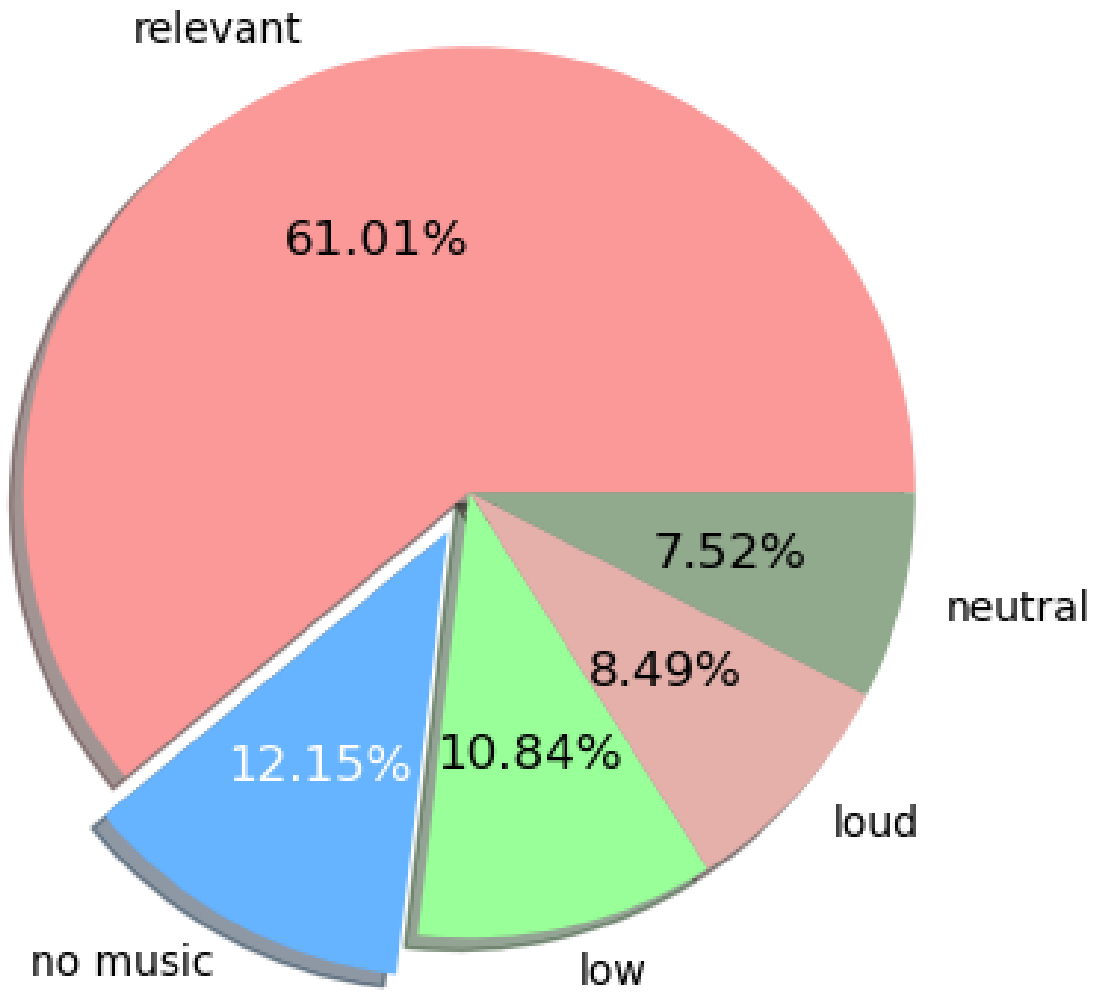
visual_first_3_seconds

- More than 34 percent of videos starts with visual of common person
- Others include anything other than humans. For example, nature, animal etc.

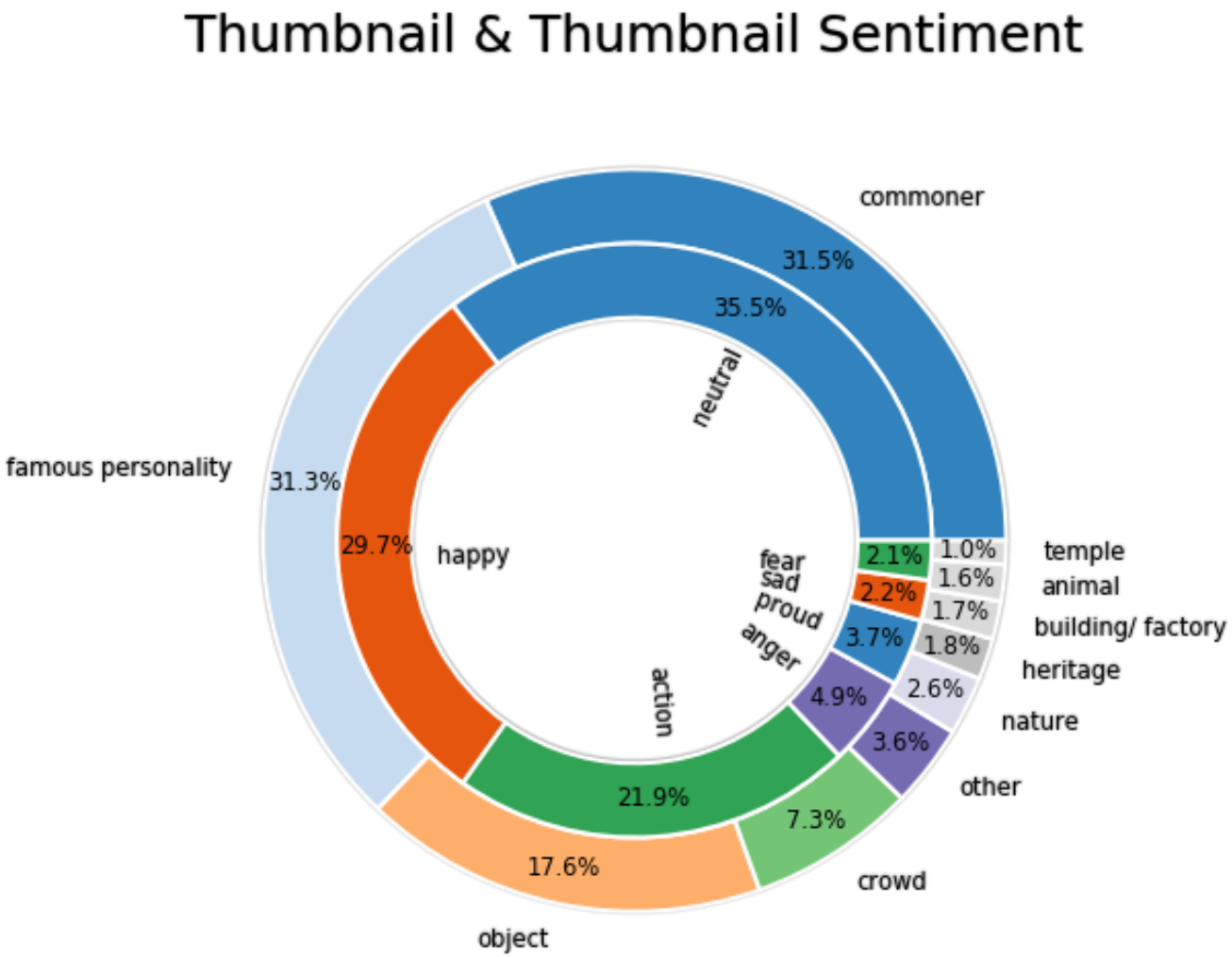


Background music type in first 3 seconds

- 12 percent of videos do not have the background music.



Thumbnail and Thumbnail Sentiment



Tags

176 videos tags were not editable

No tags were filled for 518 videos