CSC4211 - PROJECT WORK

ANALYSIS OF CLICKBAIT IN YOUTUBE VIDEOS USING ENSEMBLE MODELS

VISHAL KRISHNAN S H - 170071601142

VISHAL R K - 170071601143

GUIDED BY: MR. V. BALAJI

WHAT IS CLICKBAIT ?

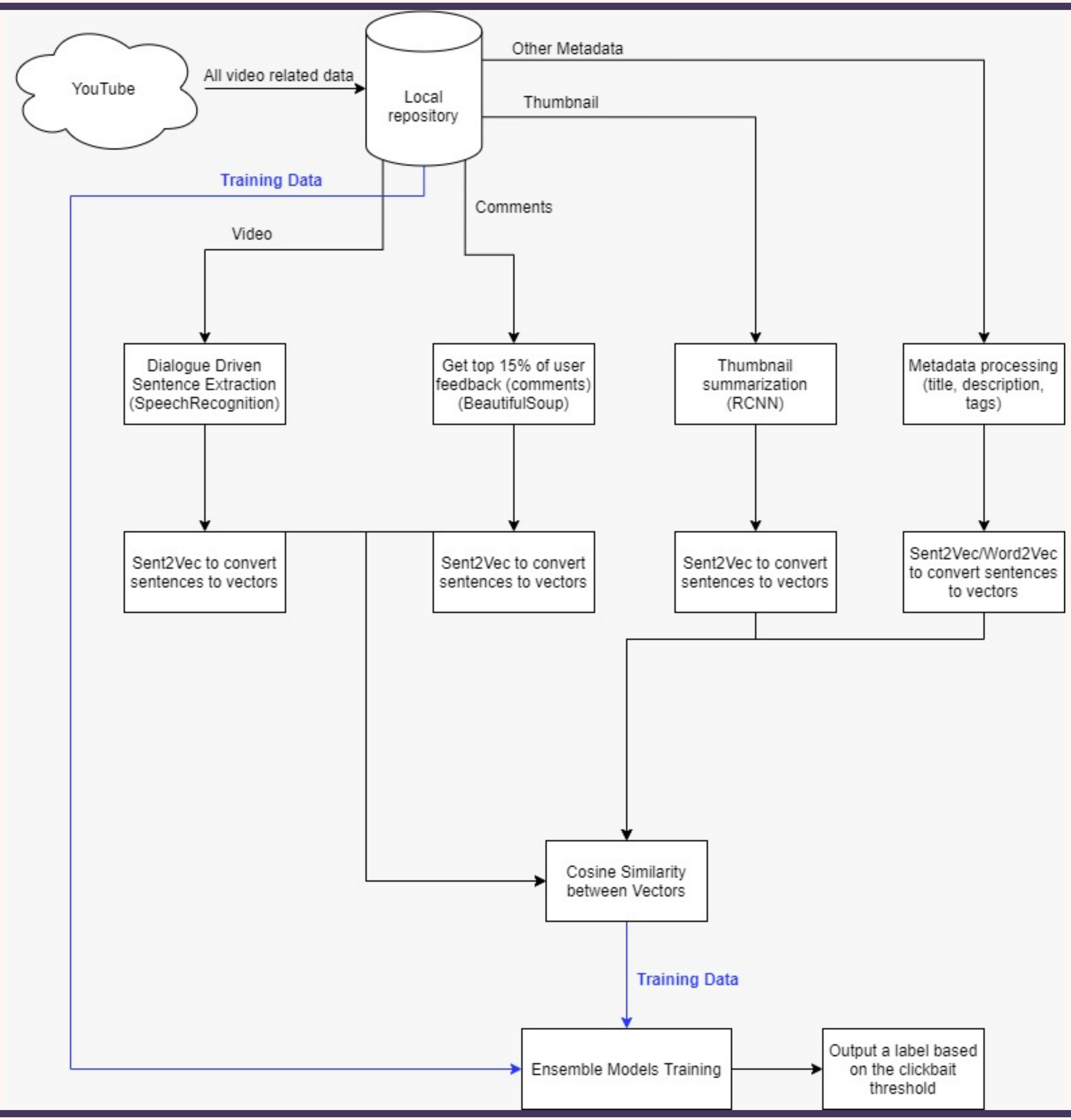
- Clickbait is a marketing device (such as a headline) designed to make readers want to click on a hyperlink especially when the link leads to content of dubious value or interest.
- Clickbait is ever prevalent in the recent world of sensationalist media at the cost of journalistic integrity.
- It has plagued the most popular video services provider on the internet - **YouTube**.
- Our project's aim is to analyze and, eventually, create a model that helps detect clickbait, not only in currently uploaded videos, but also in any future video uploaded on the platform.

SCOPE OF THE PROJECT

The scope of this project extends to achieve the following 4 goals:

- Analyze the amount clickbait present in a sample set of YouTube videos.
- Detect factors that affect the viewer's perception of what constitutes "clickbait"
- Develop a learning model that studies the aforementioned factors to categorize videos as "clickbait" and "not clickbait"
- Test it on newly, non-encountered examples to measure its effectiveness and accuracy.

PROPOSED ARCHITECTURE



ALGORITHMS USED

There are various ML and DL algorithms that will be used in our training model.

1. **RCNN** (Ensemble) – used to caption the thumbnails of videos
2. **SBERT**– used to create sentence and word embeddings (vectors)
3. Various Classification Models
   - Logistic Regression
   - Gaussian Naive Bayes
   - Decision Tree
   - Random Forest Classifier
   - K-Nearest Neighbors Classifier
   - Support Vector Classifier

FEATURES COLLECTED

Values	Data Type	Location
Creator Fed Values		
Video_ID	String	YouTube API v3
Title	String	YouTube API v3
Tags	String	YouTube API v3
Description	String	YouTube API v3
Thumbnail Caption *(generated)*	String	RCNN Ensemble Model
Video Rating	String	YouTube API v3
Video Category ID	String	YouTube API v3
Uploaded At	Timestamp	YouTube API v3
Audio Transcript	String	Audio to Speech Recognition
User Interaction Values		
Likes	Integer	YouTube API v3
Dislikes	Integer	YouTube API v3
Like-Dislike Ratio	Float	Calculated
Number of comments	Integer	YouTube API v3
Number of "Fake" Comments	Integer	Calculated
"Fake" Comment Ratio	Float	Calculated
Views	Integer	YouTube API v3
Content Creator's Data		
Channel Age	Integer	YouTube API v3
Channel Total Views	Integer	YouTube API v3
Channel Total Subscribers	Float	Calculated
Channel Made for kids	Integer	YouTube API v3
Channel number of videos	Integer	Calculated

FEATURES COLLECTED (CONTD.)

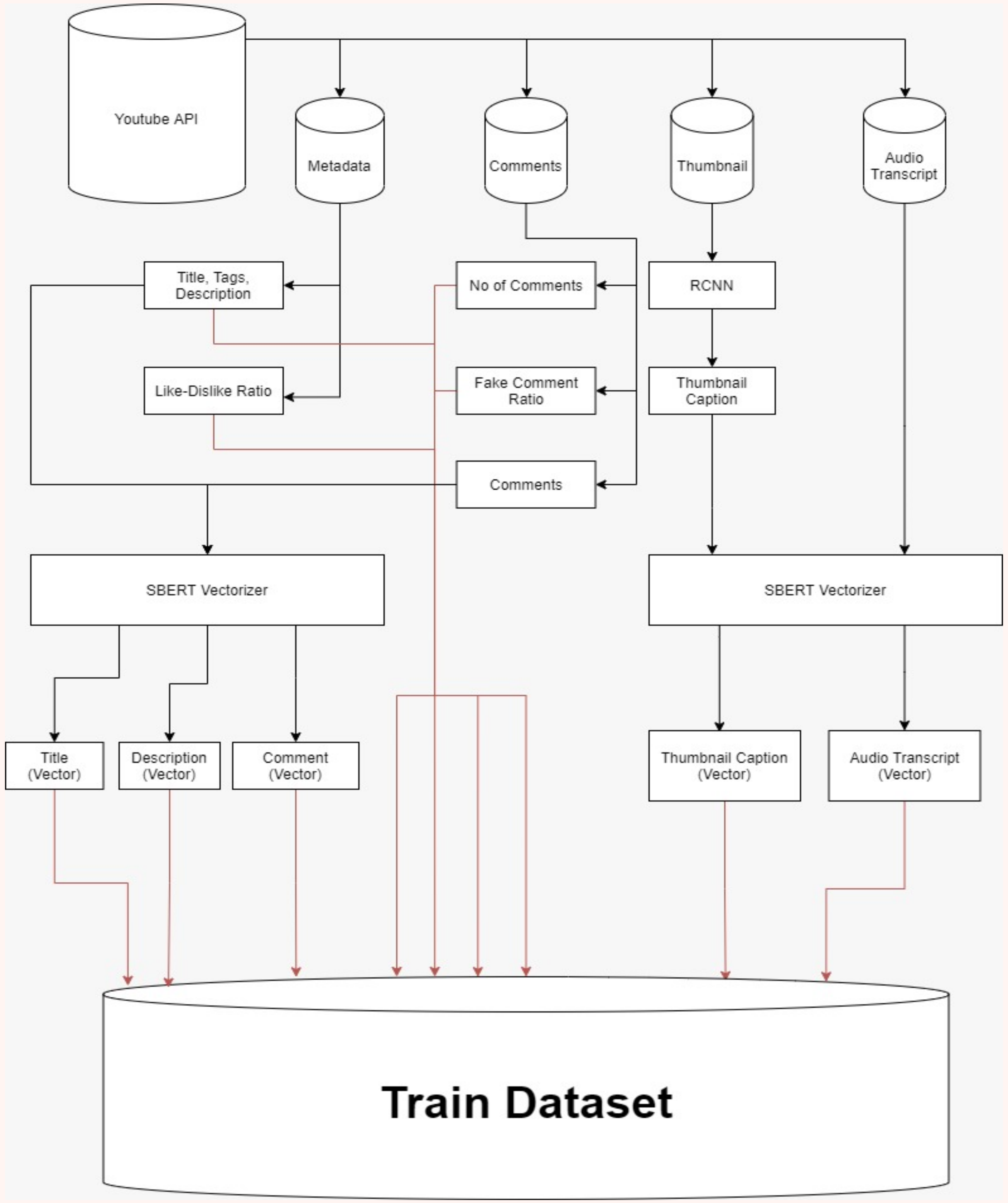
- The data surrounding YouTube videos and clickbait is very less, almost non-existent
- Manual dataset creation using:
  - Google's own YouTube API
  - BeautifulSoup
  - Youtube-Transcript API
- Data collected across various categories.

SOFTWARE AND PACKAGES

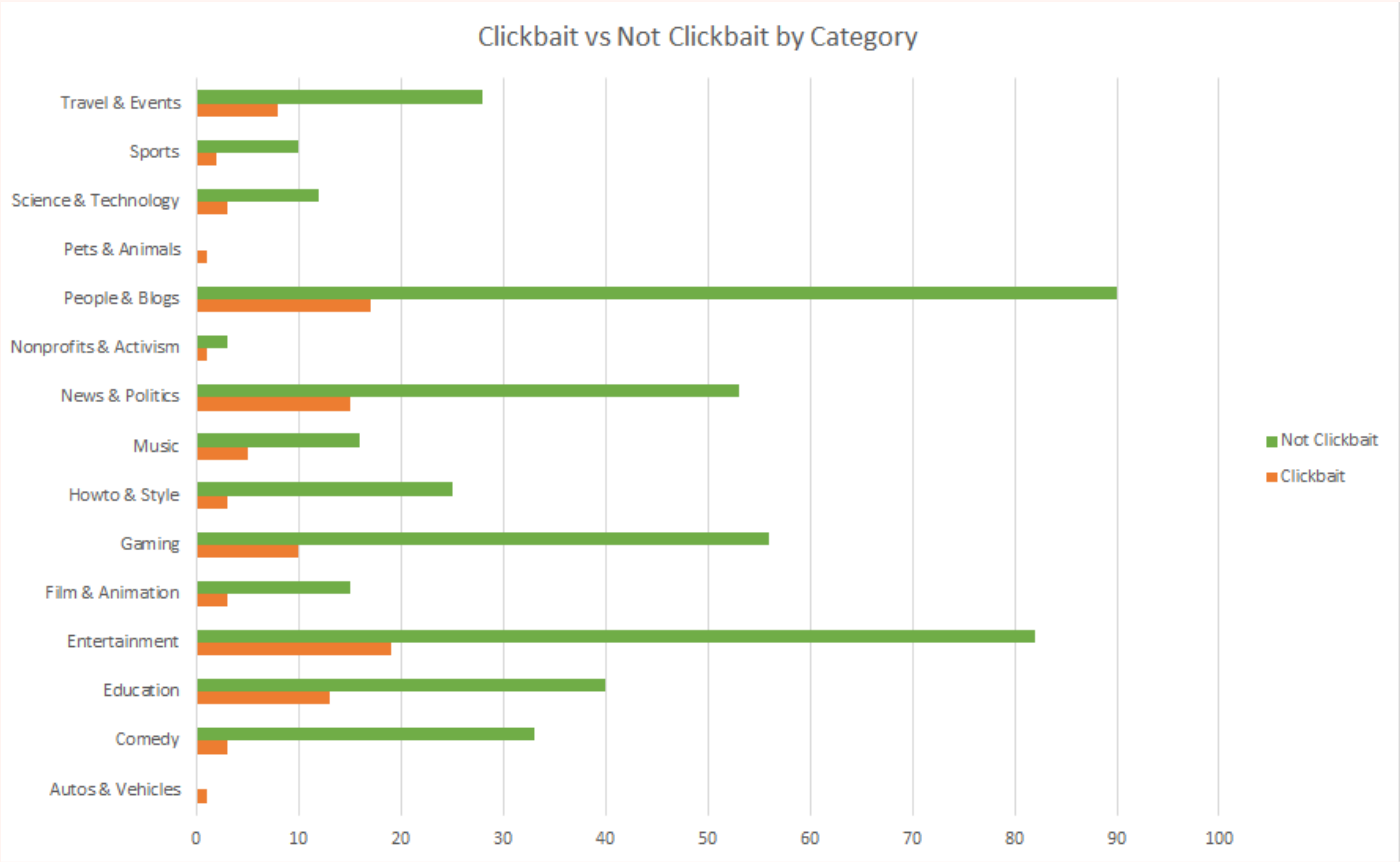
The primary tools, software, and websites used are as follows:

- Google Colaboratory
- Python 3
- YouTube API
- YouTube Transcript API
- Scikit-Learn
- Keras
- Inception V3
- NLTK
- Beautiful Soup
- SentenceTranformers

DATA EXTRACTION AND DATA PRE-PROCESSING PIPELINE



PRELIMINARY ANALYSIS ON COLLECTED DATASET



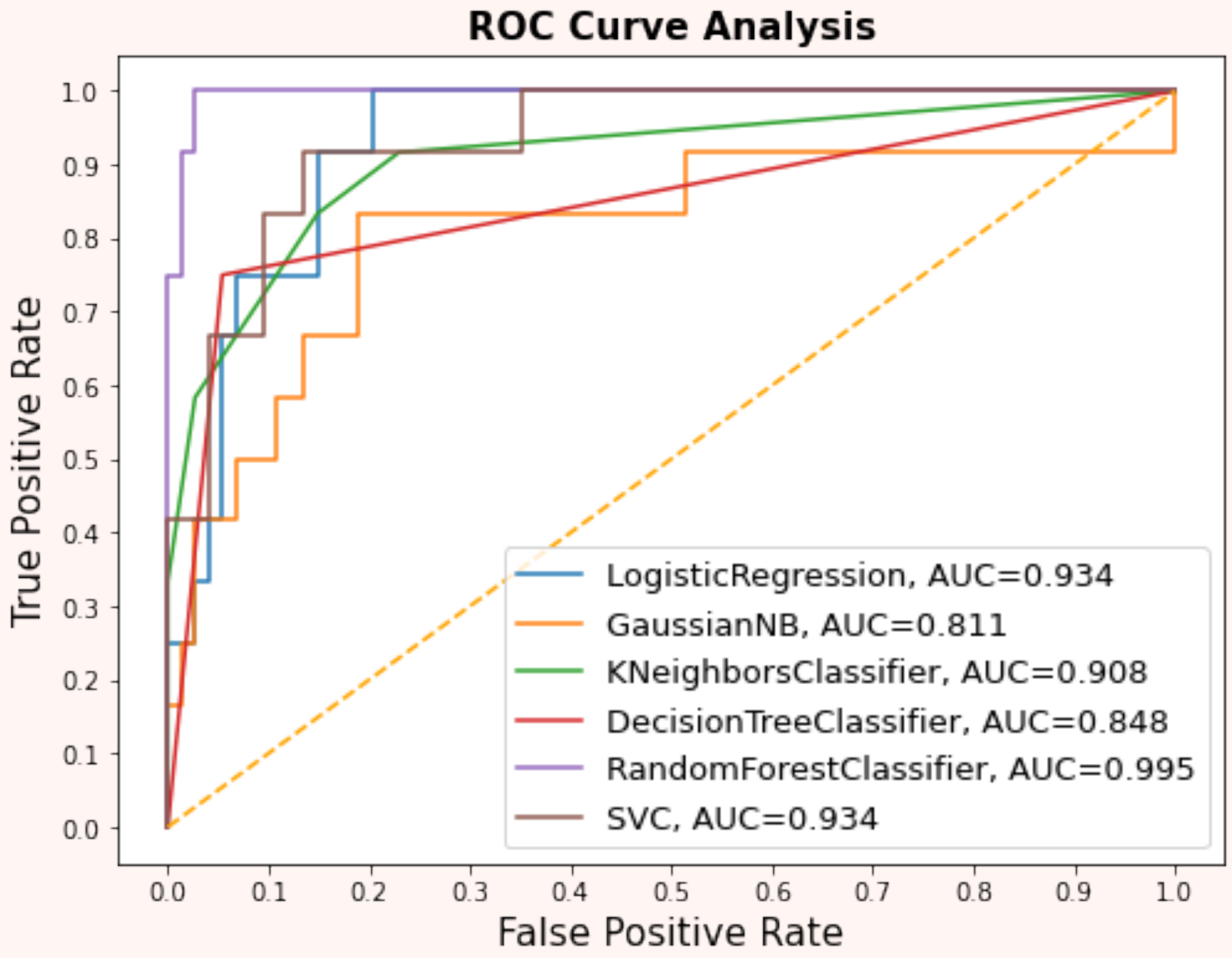
OUTPUT AND METRICS

Output - Probability percentage of the video being clickbait

Metrics used to compare classification models:

1. Precision
2. Recall
3. F1 Score
4. Accuracy
5. ROC-AUC

COMPARISON OF CLASSIFIERS



Logistic Regression

	0	1	accuracy	macro avg	weighted avg
precision	0.898734	0.571429	0.872093	0.735081	0.853064
recall	0.959459	0.333333	0.872093	0.646396	0.872093
f1-score	0.928105	0.421053	0.872093	0.674579	0.857353
support	74	12	0.872093	86	86

Gaussian Naive Bayes

	0	1	accuracy	macro avg	weighted avg
precision	0.947368	0.164179	0.337209	0.555774	0.838086
recall	0.243243	0.916667	0.337209	0.579955	0.337209
f1-score	0.387097	0.278481	0.337209	0.332789	0.371941
support	74	12	0.337209	86	86

KNN Classifier

	0	1	accuracy	macro avg	weighted avg
precision	0.935065	0.777778	0.918605	0.856421	0.913118
recall	0.972973	0.583333	0.918605	0.778153	0.918605
f1-score	0.953642	0.666667	0.918605	0.810155	0.913599
support	74	12	0.918605	86	86

Decision Tree Classifier

	0	1	accuracy	macro avg	weighted avg
precision	0.958904	0.692308	0.918605	0.825606	0.921705
recall	0.945946	0.75	0.918605	0.847973	0.918605
f1-score	0.952381	0.72	0.918605	0.83619	0.919956
support	74	12	0.918605	86	86

Random Forest Classifier

	0	1	accuracy	macro avg	weighted avg
precision	0.960526	0.9	0.953488	0.930263	0.952081
recall	0.986486	0.75	0.953488	0.868243	0.953488
f1-score	0.973333	0.818182	0.953488	0.895758	0.951684
support	74	12	0.953488	86	86

Support Vector Classifier

	0	1	accuracy	macro avg	weighted avg
precision	0.9125	0.833333	0.906977	0.872917	0.901453
recall	0.986486	0.416667	0.906977	0.701577	0.906977
f1-score	0.948052	0.555556	0.906977	0.751804	0.893285
support	74	12	0.906977	86	86

OUTPUT SCREENSHOTS

```
isclickbait('xDzcHagoYqw')

Collecting data from the video..
Title : LS | C9 vs IMT Analysis | THIS IS NOT The FUDGE FACTOR I Know...
Likes : 3623
Views : 105026
Dislikes : 55
Thumbnail :
```


```
Processing....
Our Model is 13.16% confident that this video is clickbait
```

Video 1

- Talking about a niche subject matter
- Very wordy title
- Good LD Ratio
- Model - only 13.5% confident that it is clickbait
- (thus, not clickbait)

```
isclickbait('AakARCMfu00')

Collecting data from the video..
Title : SHOCKING COMMERCIAL TRICKS WITH FOOD || Amazing Cooking Hacks
Likes : 4014
Views : 291627
Dislikes : 402
Thumbnail :
```

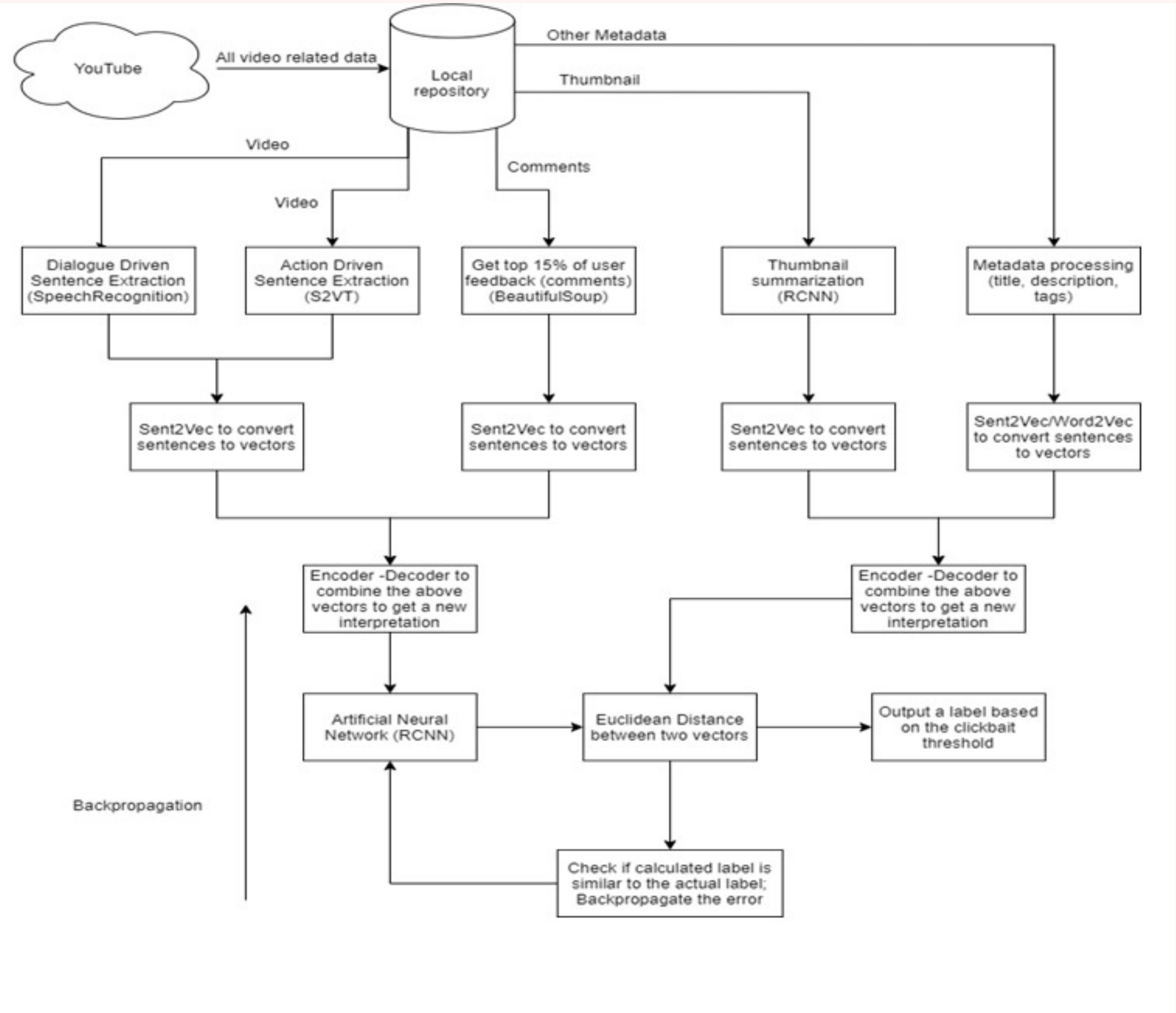
```
Processing....
Our Model is 96.30% confident that this video is clickbait
```

Video 2

- Very bright thumbnail
- Extremely catchy words
- Average LD Ratio
- Model - 96.3% confident that it is clickbait
- (thus, clickbait)

LIMITATIONS

- Biased to our opinion of what is clickbait and what is not
- Generic model for every YouTube video - very difficult
- Lack of computational resources and quality data
- Unable to implement our initial vision for this project due to the above



FUTURE ENHANCEMENTS

There are a number of enhancements that we can think of applying in the near future:

- Implementing video summarization and captioning to directly extract the essence of the video
- Building an extension that gets user feedback for the clickbait value and retrain periodically
- Consider uploader's track record of clickbait history once the model has enough data

REFERENCES

- Sarjak Chawda, Aditi Patil, Abhishek Singh, Prof. Ashwini Save ,' A Novel Approach for Clickbait Detection' - Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8
- E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages," in IEEE Access, vol. 8, pp. 61726-61740, 2020, doi: 10.1109/ACCESS.2020.2984503.
- Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, Tat-Seng Chua 'Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification' in IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 14, NO. 4, AUGUST 2012
- Andrej Karpathy,Li Fei-Fei 'Deep Visual-Semantic Alignments for Generating Image Descriptions' – Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3128-3137
- Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama,Kate Saenko, Trevor Darrell 'Long-term Recurrent Convolutional Networks for Visual Recognition and Description'
- Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan 'Show and Tell: A Neural Image Caption Generator' - Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164
- Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio 'Show, Attend and Tell: Neural Image Caption Generation with Visual Attention' - Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2048-2057, 2015.
- Subhashini Venugopalan,Marcus Rohrbach,Jeff Donahue,Raymond Mooney ,Trevor Darrell,Kate Saenko "Sequence to Sequence – Video to Text" - Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4534-4542
- Abinash Pujahari and Dilip Singh Sisodia "Clickbait Detection using Multiple Categorization Techniques"
- Shu, Kai & Wang, Suhang & Lee, Dongwon & Liu, Huan. "(2020). Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements" 10.1007/978-3-030-42699-6_1.

THANK YOU