Exploring the Indian Political YouTube Landscape: A Multimodal Multi-Task Approach

Adwita Arora[1], Naman Dhingra[1], Divya Chaudhary[2], Ian Gorton[2] and
Bijendra Kumar[1]

[1]*Netaji Subhas University of Technology, New Delhi, India*
[2]*Northeastern University, Boston, MA 02115, United States*

Abstract

Social media profoundly influences all facets of our lives, including politics. Political parties, politicians, and media outlets have strategically cultivated their social media presence to engage with the public. However, with the advent of freely available Internet services in India, there has been a rising proliferation in the community of independent content creators on YouTube, with many getting millions of views per video. In this study, we present a novel multimodal dataset of videos, taken from 20 independent and influential content creators, annotated for five socially and politically relevant labels with a high inter-annotator score (0.820 - 0.956 Cohen's Kappa Score) falling under the categories - Humour/Satire, Opposition/Criticism, Support/Advocacy, and Informational/Analysis. We consider three modalities in our dataset - textual (title and description of the video), visual (thumbnail) and audio (MFCC coefficients and additional spectral and temporal features) modalities. We also perform preliminary classification on our dataset using an early fusion multimodal model, combining audio, visual and textual modalities, which performs better than other unimodal and bimodal approaches, yielding a Macro-F1 score of 0.8742 and ROC-AUC score of 0.769. By introducing this novel dataset, we aim to stimulate further investigation within the domains of opinion dissemination across social networks and the analysis of multimodal content, especially within the Indian context.

1. Introduction

YouTube is the most popular video-sharing platform, created in 2005, with over 2 billion monthly active users. YouTube's popularity has witnessed a massive spike across the globe in the past decade, especially in India, where it is one of the most visited websites with millions of monthly users. Videos on YouTube can be uploaded on a variety of topics, including but not limited to sports, education, entertainment, news, music and gaming. The viewers of these videos can also 'like' them and leave their thoughts as comments. The tremendous amount of information available on YouTube in the form of videos or comments makes it a favoured source for conducting research.

India, being the world's largest democracy, has a rich and vibrant political and social history. Politics is an essential aspect of the lives of Indians, making it one of the country's most deliberated and debated topics. With the advent of social media, people are voicing their opinions and concerns in a manner that has never been more convenient. This could be through tweets on Twitter, consuming or creating content on YouTube, or posting on Facebook, among many other such avenues. Political parties and politicians also maintain social media profiles to connect and engage with citizens.

With the popularity of YouTube in India and the importance of politics in Indian society, analyzing the content put out by Indian creators on politics, as well as the response of the audience to it becomes essential. This research aims to analyze politically and socially relevant videos uploaded by independent content creators on YouTube. For this study, we manually chose 20 prominent and diverse YouTubers regularly making content on the politics and society of India. We selected the 20 most viewed relevant videos of each YouTuber to form our dataset.

Videos uploaded to YouTube rarely subscribe to a single topic, given their detailed nature. Especially with political videos, creators can employ different communication techniques that strategically capture the audience's attention and convey their point across. Videos can differ within the stance taken by the creator as well as what justification they provide towards it. This makes it important to model the understanding of these videos on a multi-task basis, where a single video can have multiple labels.

Therefore, for each video, we annotate it in five categories - the presence or absence of humour/satire/irony, support or opposition of any political entity, and whether the video is a fact-based analysis or personal opinion. Each of these categories is independent of the other, and the presence of one category does not affect another. We draw insights on the data collected using topic modelling on comments and keyphrase analysis on titles as well as text extracted from the thumbnails.

The final portion of this paper is devoted to a multi-task, multimodal classification of the dataset. Each video can be represented as a combination of three modalities - audio, visual and textual. After extracting relevant features from them, we employ an early fusion classification model on these modalities.

YouTube has been a source of ample classification and analysis tasks in the past due to the massive volume of public data available on various topics. Apart from the videos, comments also act as a rich source for analysis as they serve as responses from the audience, both of which can be studied individually or in unison. Kang et al. [1] analyzed the Mukbang-related content on YouTube along with news and observed how behaviours like overeating are linked directly to the video's popularity, while Papadamou et al. [2] analyzed the Incel community and the abundance of toxic and misogynistic comments on these forums.

Works like [3] and [4] are focused on the analysis of YouTube comments for exposure of children to inappropriate content and transphobic/homophobic content identification, respectively. Latorre and Amores [5] presented a topic modelling analysis of xenophobic and racist comments in Spanish directed at migrants and refugees.

A single video comprises many modalities, and several methods of classification and analysis have been developed to handle these modalities. Yousaf and Nawaz [6] proposed a novel EfficientNet-BiLSTM approach for detecting inappropriate content in animated cartoon videos targeted at children. They extracted video descriptions using EfficientNet, a pre-trained CNN

model, which was then fed to BiLSTM to learn representations. They showcased how deep learning methods perform better than traditional machine learning approaches. [7] and [8] proposed techniques that deal with internet memes, which refer to the image + textual modality. Chauhan et al. [9] extended the M2H2 data by adding parallel English translations and annotating each entry for emotion and sentiment classes. They also proposed a multimodal multitask classification using a context transformer with sentiment and emotion embeddings baseline. They displayed that combining all three modalities led to the best results.

The creation of a well-annotated dataset is the backbone of any systematic research. Shahi [10] presented a semi-automated annotation framework for multilingual, multimodal social media data. Expertly annotated multimodal and multilabel datasets have also been proposed on diverse subjects. Chauhan et al. [9] and Christ et al. [11] proposed datasets on humour detection where M2H2 was annotated for numerous occurrences from a well-known Hindi TV show, and Passau-SFCH was annotated for humour along the sentiment (Positive or Negative) and direction (towards self or towards others) dimensions, respectively. Gupta et al. [12] presented 3MASSIV, a dataset of about 50,000 expertly annotated multilingual short videos from a sharing platform called Moj. Other works like Khan et al. [13] present Vyaktitv - a multimodal dataset consisting of participants' audio and visual recordings and their Hinglish transcriptions for personality detection.

To our current understanding, we encountered a challenge in locating a multimodal dataset that has been annotated to encompass five distinct socio-political labels such as ours.

2. Dataset

Details of the YouTubers selected for this study are given in Table 1.

To build the dataset, we first prepared a list of popular Indian YouTubers creating content on socially and politically relevant topics. We manually selected 20 YouTubers, each with a subscriber count of over 330,000, as of February 2023. We then selected 20 of the most viewed videos from each of these YouTubers, post a manual removal of those videos that were either irrelevant to politics or had a duration greater than 25 minutes or less than 5 minutes. This was done in order to ensure consistency among all videos. We then used the YouTube API to collect the *title* of the video, the *description* of the video, the *number of likes* received and *view count* as of February 2023 and the *thumbnail* for each of the 400 videos. [1] Details of the dataset are given in Table 1. The audio of the videos was downloaded as a .mp4 file using the pyTube Python library[2], later processed using the Librosa library[3].

---

[1] Videos uploaded to YouTube fall under its "fair use" guidelines, which is a legal doctrine that says the use of copyright-protected material under certain circumstances is allowed without permission from the copyright holder. In the United States and India works of research may be considered fair use if done fairly (https://support.google.com/youtube/answer/9783148?hl=en).

[2] https://github.com/pytube/pytube

[3] https://github.com/librosa/librosa

YouTuber	Avg. View Count	Avg. Like Count	Avg. Comment Count	Avg. Subscriber Count
Abhisar Sharma	1436644.15	73404.75	4551.5	1970000
The Jaipur Dialogues	521948.3	22276.6	3379.55	952000
String	864882.75	93449.35	15232.95	1140000
Kumar Shyam	293947.35	14661.55	2574.05	396000
Soch by Mohak Mangal	819061.9	51425.15	7932.7	2210000
Sushant Sinha	1160897.4	48535.85	6623.85	1210000
The Deshbhakt	2476270.2	135741.1	12644.1	3000000
Kroordarshan	123431.6	8621.85	471.15	384000
Punya Prasun Bajpai	1361489.9	48938.15	4535.65	2720000
Sarthak Goswami	429011.15	26884.85	1807.35	658000
Sakshi Joshi	873971.75	31364.55	3318.05	826000
Open Letter	153227.1	15980.9	2452.9	420000
Harsh Vardhan Tripathi	281155.05	11683.15	1062.75	383000
Dhruv Rathee	5889225.4	382734.3	45536.1	10700000
AKTK	1315234.55	44199.55	6153.45	1260000
Being Honest	1006400.2	73883.05	4431.2	962000
Ajit Anjum	2550387.55	58477.5	6182.15	3580000
The Manish Thakur Show	307334.85	11899.4	920.5	331000
The Sham Sharma Show	953314.55	86985.85	10655.4	1060000
DO Politics	666685.8	65583.3	8249.85	713000
Overall Dataset	**1174226.075**	**65301.2375**	**7433.62**	**1743750.0**

Table 1
Dataset Details
Values averaged over 20 videos

2.1. Annotations

In recent years, the use of NLP and ML techniques to study politics has drawn more and more interest from the research community. Our goal is to promote significant research improvements across computational and socio-political areas, specifically in the Indian context, using this multi-task framework and recognising the multimodal complexity of the data.

We employed two undergraduate students with a strong grip on both Hindi and English to annotate the videos. We considered five tasks for this annotation process - Humour/Satire, Opposition/Criticism, Support/Advocacy, and Informational/Analysis.

We have based our task definitions around a political entity i.e.: directed at or in reference to. We define a **political entity** as a politician, a political party, the supporters of a political party, and **political event** as any event that is linked to recent socio-political issues. Each of the videos was then annotated based on the following task definitions which were given to both the annotators along with an example. Each video was annotated individually for each task. Examples for each task are given in Figure 1.

- **Task 1: Humour/Satire** Over the past few years, researchers have become increasingly interested in the topic of humour and satire detection on its own. When viewed from a political angle, humour detection is a potent indicator that can be used to determine how

(a) Humour/Satire
Title - "Sambit Patra: Best moments from the motormouth BJP spokesperson! | The DeshBhakt with Akash Banerjee"
Audio - *"या फिर देखिए, एक सिंगर, par-excellance! अरे किशोरे भी शर्मा जाए! "*
Trans. - *"He is a singer par-excellance, even better than Kishore!"*

(b) Opposition/Criticism
Title - मीडिया के गिद्धों से मिलिए}.... Godi Media| Suresh Chavhanke | Amish Devgan | ABP news |
Audio - "...लेकिन जब ऐसे लोग, एह सिर्फ सुरेश चभांके की बात नही है | Your very dear अंबानी का चैनल दिन भर नफरत परोसता है..."
Trans. - *"...this is not only about Suresh Chavhanke, your very dear Ambani's channel spews hate all day..."*

(c) Support/Advocacy
Title - "Supreme Court \hindifont{से बड़ी खबर, PM Modi को हरानेवाला मिल गया|} Sushant Sinha | Pegasus | Punjab"
Audio - "...जनता की उम्मेदे नरेंद्र मोदी से यहाँ है, सात्वे आस्मन पर, और नरेंद्र मोदी उन उम्मेदों को पुरा करने के लिए 24 घन्टे सातो दिन काम कर रहे है ..."
Trans. - *"... the public has very high hopes from Narendra Modi, and Narendra Modi is working 24 hours a day 7 days a week to fulfil those hopes ..."*

(d) Informational/Analysis
Title - "Why Narendra Modi is so popular" - Soch by Mohak Mangal
Audio - "...इसिलिए कयी political journalists and experts ने यह दावा किया है की प्रधान मन्त्री मोदी ने एक 'teflon' इमेज develop कर्ली है ..."
Trans. - *"... this is why many many political journalists and experts have claimed that PM Modi has developed a "teflon" image ..."*

(e) Opinion
Title - "The Problem with Arnab Goswami | Opinion by Dhruv Rathee"
Audio - "...और यह noisy journalist है, लेकिन यह मेरी राय है यहाँ पर..."
Trans. - *"...and he is a noisy journalist, but this is my personal opinion..."*

Figure 1: Examples for each of the labels

the general public feels about a situation or an entity. Politicians and content providers alike can use it as a tactic to interact with and draw the public's attention. More nuanced issues like misinformation and manipulation can also be masked as humour, which needs to be addressed. For the video to fall under this category, there is at least one mention

Label	Cohen's Kappa Score
Humour/Satire	0.956
Opposition/Criticism	0.934
Support/Advocacy	0.820
Informational/Analysis	0.857
Opinion	0.874

Table 2
Cohen's Kappa Score for each label

of a joke, meme, caricature or satirical/sarcastic comment directed at a political entity or with regards to a political event, either uttered by the creator, visible in the video or present in the title or description of the video.

- **Task 2: Opposition/Criticism** In this category, the aim is to detect any explicit opposition or criticism of a political entity or the actions of the political entity with regard to a political event. This could take place either as an utterance, demonstrated visually in the video or textually in the title or description of the video. A common example is the criticism of "Godi Media" or pro-government media houses for not focusing on important issues [14]. Analysis of opposing or critical material is an important factor that sheds light on public opinion, highlighting the degree of disagreement and ideological differences. Issues receiving the most opposition and criticism tend to be those that are most relevant to the public.

- **Task 3: Support/Advocacy** Similar to the Opposition/Criticism category, the aim here is to detect any explicit support or advocacy of a political entity or the actions of the political entity with regard to a political event. This could take place either as an utterance, demonstrated visually in the video or textually in the title or description of the video. For example, the creator could endorse policies introduced by the incumbent politicians. We expect that studying Task 2 and Task 3 in unison also offers useful insights with respect to public opinion fluctuation towards entities and events as well as bias detection.

- **Task 4: Informational/Analysis** If the nature of the video is informational or is an analysis of a relevant political entity or event, where the YouTuber explains the events to the audience using suitable sources (news articles, scholarly publications or government documents), the video will be informational in nature.

- **Task 5: Opinion** A video will fall under this category if the YouTuber voices their personal opinion on any political entity or event, with or without justification. Detecting when content is an opinion piece vs. factual information can be used in the downstream modelling task of misinformation detection. This task also has uses in detecting deviation of public opinion from reality as well as in combating confirmation bias.

To measure the quality of our annotations, we choose Cohen's Kappa statistic [15] which is a measure of the Inter-Annotator Agreement (IAA). The scores obtained for each label are mentioned in Table 2 and show the presence of high agreement for each of the labels.

Label	Number of occurrences
Humour/Satire	143
Opposition/Criticism	209
Support/Advocacy	105
Informational/Analysis	301
Opinion	319

Table 3
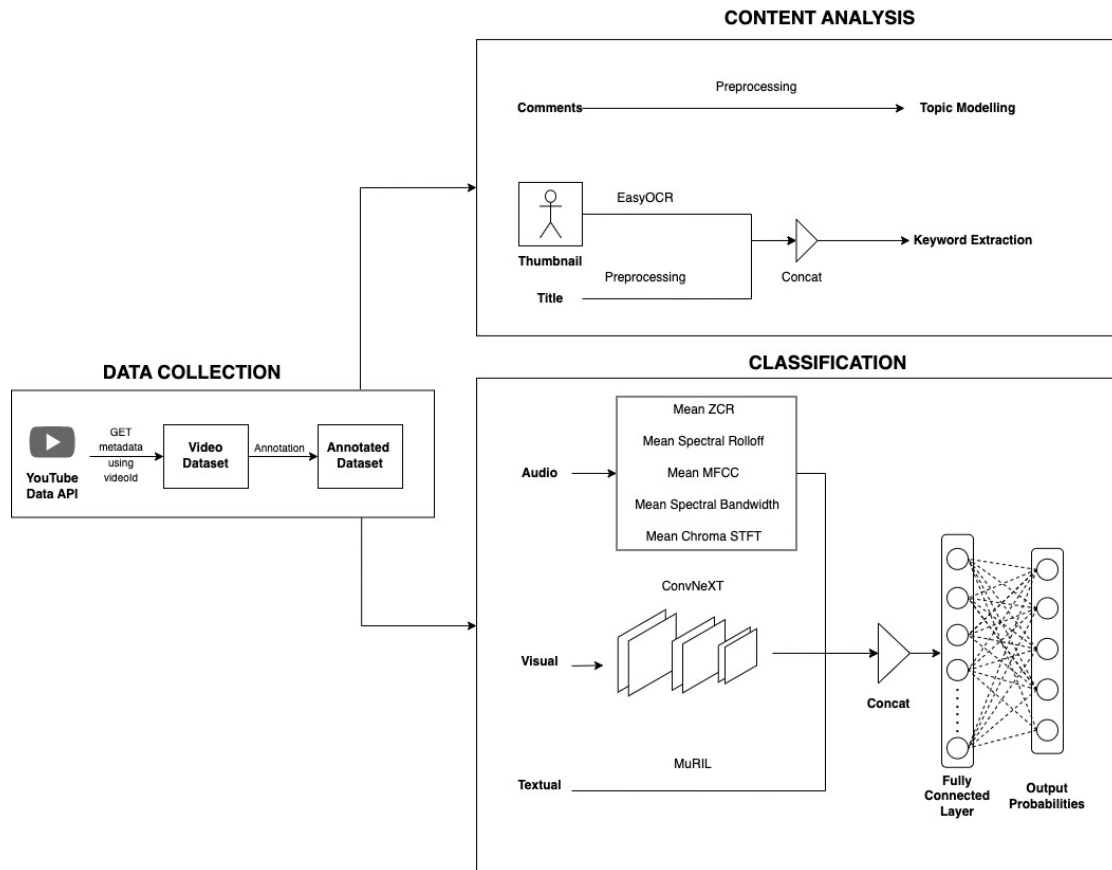Number of occurrences for each label

Figure 2: Complete Pipeline

3. Methodology

Political videos uploaded to YouTube form interesting inputs to both analysis as well as classification studies. For our research, we have chosen to focus on both of these major tasks:

1. **Content Analysis** By performing a thorough analysis of the different metadata features extracted using the YouTube Data API, we want to understand both the kind of content political YouTubers choose to put out as well as the response of the audience to it.

2. **Classification** We also provide different experiments on the multitask, and multimodal classification of videos into the five labels as described above.

Figure 2 shows the complete pipeline for the content analysis and classification tasks.

3.1. Content Analysis

Each feature collected from a video offers significant insight into the content of the video. For example, the title and the thumbnail are often designed to captivate the audience's attention, since they are the first features spotted. This has set a trend of content creators using "clickbait" to mislead viewers, leading to numerous studies on clickbait detection [16, 17, 18]. On the other hand, comments act as an outlet for the audience's reaction to the video. Other statistical features, such as the number of likes and view count, indicate the acceptance and virality of the videos, respectively. We used topic modelling and keyword extraction to analyze these features.

3.1.1. Topic Modelling

An unsupervised method for identifying the most meaningful topics from a given corpus of text is called topic modelling. BERTopic [19], Top2Vec [20] and Latent Dirichlet Allocation (LDA) [21] are popular topic modelling methods. We have used BERTopic for our study, an approach that uses a class-based implementation of the TF-IDF method to cluster embeddings obtained from pre-trained transformers to produce pertinent topics. BERTopic was especially favoured for its multilingual support [22]. To perform topic modelling on the comments, we first extracted the 25 most "relevant" comments along with their timestamps from each of the 400 videos resulting in a corpus of 10,000. BERTopic provides a convenient function to visualize the topics generated with time. Figure 3 depicts this visualization. One of the biggest spikes was found in topic 11 in January 2021, which would allude to the farmers' protest on Republic Day, leading to a massive nationwide debate [4]. Other topics include religion (*"hindu", "muslim"*), political parties (*"congress", "bjp"*) and specific events (*"farmer", "election"*).
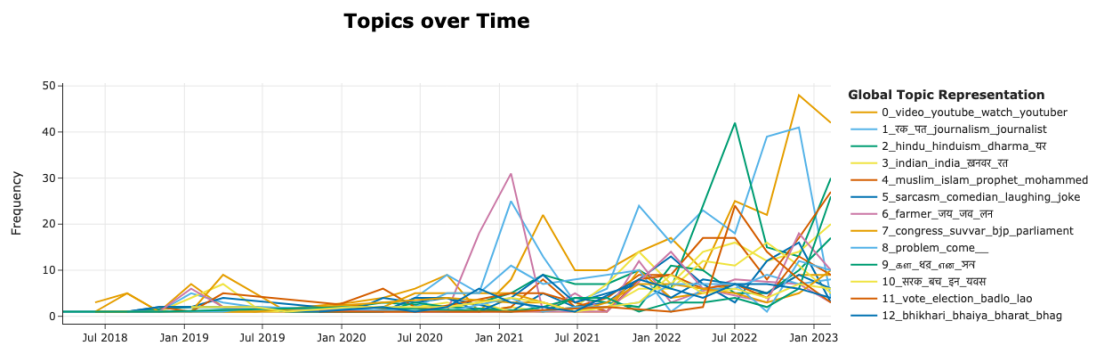

Figure 3: Time-based topic analysis on comments using BERTopic

---

[4]https://en.wikipedia.org/wiki/2021_Indian_farmers%27_Republic_Day_protest

3.1.2. Keyword Analysis

Keyword	Relevance Score
'modi'	0.00464
'godi'	0.00750
'news'	0.01084
'media'	0.01178
'bjp'	0.01252
'week'	0.01315
'episode'	0.01422
'noise'	0.01717
'india'	0.01730
'top'	0.01923
'show'	0.01946
'kumar'	0.01992
'adani'	0.02055
'views'	0.02089
'badi*'	0.02180
'rahul'	0.02392
'explained'	0.02448
'BJP*'	0.02449
'gandhi'	0.02586
'narendra'	0.02723

Table 4
Top 20 most relevant keywords and corresponding relevance scores given by YAKE!
Translated from Hindi

Extraction of 'keywords' or 'keyphrases' from a document is a method for the succinct representations of its content. Keyword extraction is widely used in research areas like opinion mining, information retrieval systems, document clustering, and other NLP tasks [23]. Many popular approaches like YAKE! [24, 25], RAKE [26] and KeyBERT [27] are used for keyword extraction. YAKE!, being a keyword extraction method based on statistical text features, is domain-independent and language-independent, thus being the ideal choice for our study. YAKE! returns a list of keywords along with a relevance score; the lower the score, the more relevant the keyword is to the document. To perform keyword extraction using YAKE!, we concatenated the title string and the text extracted from the thumbnail using the EasyOCR Python library[5]. The concatenated strings were preprocessed by converting them to lowercase and removing Hindi and English stopwords using the NLTK Python library[6]. We found the names of some YouTubers occurring in some of the strings, which we removed, to keep them from appearing as keywords. Table 4 shows the top 20 keywords extracted from the strings, along with their relevance scores.

---

[5]https://github.com/JaidedAI/EasyOCR
[6]https://www.nltk.org

3.2. Classification

For a given video, three modalities were extracted, namely audio($\mathbb{A}$), text($\mathbb{T}$) and image($\mathbb{V}$). Given these modalities, $\mathbb{A}$, $\mathbb{T}$, $\mathbb{V}$ our task is to predict the binary values for each of the five labels.

3.2.1. Feature Extraction

1. **Text** The text modality for each video is a combination of its title and description. We first concatenate the title of the video with its description. We preprocess the concatenated string to remove whitespaces, punctuations and URLs. In our dataset, the text strings were either in English, Hindi, Romanized Hindi or a combination of them. To extract the features from the text in these languages, we utilised MuRIL [28]. MuRIL, or Multilingual Representations for Indian Languages, is a language model built for 16 Indian languages and English. MuRIL has been shown to outperform the pre-existing multilingual models, such as mBERT [29], on many NLP tasks for Indian languages. For each text string $t \in \mathbb{T}$, the obtained feature embedding is a $\mathbb{R}^{1 \times d_t}$ vector. The text embeddings for the entire dataset is a $\mathbb{R}^{n \times d_t}$ vector where d_t is 768.

2. **Image** The image modality for each video is the thumbnail. We use ConvNext [30], a purely convolutional vision processing model, to extract the embedding for each thumbnail. We use the ConvNext-T model pre-trained on the ImageNet-1k dataset. For each thumbnail $\mathbb{I} \in \mathbb{V}$, ConvNext returns a $\mathbb{R}^{1 \times d_i \times 7 \times 7}$. The image embeddings for the entire dataset is a $\mathbb{R}^{n \times d_i \times 7 \times 7}$ vector. This was reduced to a $\mathbb{R}^{n \times d_i}$ vector using a max pooling and a flatten layer, where d_i is 768.

3. **Audio** To represent the audio features of each video we extract the following features using the Librosa library:

   - **Mel Frequency Cepstral Coefficients (MFCC)**[7] are one of the most frequently used audio characteristics. This feature is produced by performing a cosine transformation on the power spectrum's logarithm, which is translated onto the mel scale as evenly spaced frequency bands. The Mel scale is based on the characteristics of the human auditory system, which is better able to discern between sounds that are represented on this scale.

   - **Chroma STFT** is produced using a Fast Fourier Transform (FFT) on the audio and a series of filters to transform the power spectrum into a chromatic scale.

   - **Spectral Centroid** is taken from each frame of a magnitude spectrogram after being normalized.

   - **Spectral Bandwidth** is the width of the audio signal's power spectrum as measured at a specific level below the peak frequency.

   - **Rolloff** represents the frequency below which a certain percentage (85% by default) of the signal's total spectral energy is contained for each frame.

   - **Zero Crossing Rate** is a measurement of the audio signal's frequency content that shows how many times per second the audio signal crosses the zero axis.

---

[7]https://librosa.org/doc/latest/generated/librosa.feature.mfcc.html#librosa.feature.mfcc

Modalities	Representation	Macro-F1	ROC-AUC
Text only	MuRIL	0.326	0.521
Image only	ConvNeXT	0.740	0.705
Audio only	MFCC*	0.478	0.5
Text + Image	MuRIL + ConvNeXT	0.755	0.759
Text + Audio	MuRIL + MCFF*	0.503	0.5
Audio + Image	MCFF* + ConvNeXT	0.785	0.768
Text + Audio + Image	MuRIL + MCFF* + ConvNeXT	**0.8742**	**0.769**

Table 5
Macro-F1 and AUC Score for all modalities
Mean values of MCFF, Spectral Rolloff, Chroma STFT, Spectral Bandwith and Zero Crossing Rate

The collected features for audio thus consist of 20 values of the MFCCs along with 5 values comprising the mean Chroma STFT, Spectral Centroid, Spectral Bandwidth, Rolloff and Zero Crossing Rate. For each audio $a \in \mathbb{A}$, the obtained features is a $\mathbb{R}^{1 \times d_a}$ vector. The overall audio features thus form a $\mathbb{R}^{n \times d_a}$ vector where d_a is 25.

4. Results and Discussion

Our multimodal classification is an early fusion model, where we concatenate the embeddings received from each of the three modalities and feed it to a fully connected neural network of three layers. We kept the train-test split at 80%. The loss function chosen to minimize was the Binary Cross Entropy Loss *(BCELoss)*. We also perform ablation studies on different combinations of these modalities, namely unimodal ($\mathbb{A}$, $\mathbb{V}$, $\mathbb{T}$) and bimodal ($\mathbb{A}+\mathbb{T}$, $\mathbb{A}+\mathbb{V}$, $\mathbb{T}+\mathbb{V}$). The metrics computed using macro-averaging are provided in Table 5.

Among all the unimodal and bimodal models, it has been noted that the presence of image-based features yields the highest Macro-F1 and ROC-AUC scores. This observation finds support in the practice of YouTubers who strategically model thumbnails and titles to not only offer a glimpse of the video's content but also to effectively draw the audience in.

The early fusion tri-modal model of $\mathbb{A}+\mathbb{T}+\mathbb{V}$ performs the best out of all combinations of modalities, with a Macro-F1 score of 0.8742 and ROC-AUC score of 0.769.

5. Conclusions

The consumption of content on social media sites, like YouTube, has grown manifold over the last decade. This has led to an exponential rise in the creation of short-form and long-form content on various topics, ranging from comedic videos to documentaries. In this study, we analyzed one such content creation topic, political videos uploaded by independent Indian content creators. We annotated around 400 videos collected from YouTube for different socially and politically relevant labels. We performed a content analysis on our annotated dataset using BERTopic for topic modelling and YAKE! for keyword extraction. We also applied an early fusion multimodal model on the features extracted using state-of-the-art backbone representations,

namely MuRIL for text, ConvNeXT for images, and MCFF, ZCR, Spectral Bandwidth, Chroma STFT and Spectral Rolloff for audio. Our classification model yielded a Macro-F1 score of 0.8742. Compared to other unimodal and bimodal models, the early fusion model yielded significantly better results.

6. Future Work

Future work that focuses on a number of important areas of development will raise the calibre and scope of this research. Here are some directions we want to go in:

1. **Experimentation with other fusion models** In this paper, we used an early fusion model, combining modalities before classification. However, there are alternative fusion techniques that warrant exploration, such as late fusion models, where each modality is processed independently before being integrated with other modalities, and attention-based fusion models where the importance of different modalities is assessed with respect to the task at hand, or ensemble models, which combines the strengths of multiple prediction models to improve results.

2. **Audio feature extraction** The features extracted for audio in this study are numerical metrics that regrettably fail to capture the nuances of speech, especially code-mixed Hindi-English speech, which is a predominant mode of communication in India. Experimenting with other audio feature extraction methods, for example, using transcripts to capture semantic meaning, Mel-frequency spectrograms to capture phonetic variation or transformer-based models that are distinguished for their contextual understanding can offer more sophisticated results.

3. **Extending the dataset** We chose to annotate data collected for five tasks for the purposes of this study. However, the methods of classification and analysis can be extended to include even more relevant labels that cover more NLP and discourse analysis tasks. This includes the detection of hate speech towards marginalised communities veiled as opinions, misinformation and fake news detection or the spread and polarization of public opinions over time.

4. **Multilingual and cross-regional support** While our selection procedure primarily focused on YouTube channels that offered content in Hindi or English, it's important to recognise that a more inclusive approach is necessary for a thorough representation of India's political environment. To adequately capture the complex and varied political narratives that arise across the nation's various linguistic and cultural realms, region-specific content must be included.

Acknowledgments

References

[1] E. Kang, J. Lee, K. H. Kim, Y. H. Yun, The popularity of eating broadcast: Content analysis of "mukbang" YouTube videos, media coverage, and the health impact of "mukbang" on public, Health Informatics Journal 26 (2020) 2237–2248. URL: http://journals.sagepub.com/doi/10.1177/1460458220901360. doi:10.1177/1460458220901360.

[2] K. Papadamou, S. Zannettou, J. Blackburn, E. De Cristofaro, G. Stringhini, M. Sirivianos, "How over is it?" Understanding the Incel Community on YouTube, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–25. URL: https://dl.acm.org/doi/10.1145/3479556. doi:10.1145/3479556.

[3] S. Alshamrani, A. Abusnaina, M. Abuhamad, D. Nyang, D. Mohaisen, Hate, Obscenity, and Insults: Measuring the Exposure of Children to Inappropriate Comments in YouTube, in: Companion Proceedings of the Web Conference 2021, WWW '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 508–515. URL: https://doi.org/10.1145/3442442.3452314. doi:10.1145/3442442.3452314.

[4] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments, 2021. URL: http://arxiv.org/abs/2109.00227, arXiv:2109.00227 [cs].

[5] J. P. Latorre, J. J. Amores, Topic modelling of racist and xenophobic YouTube comments. Analyzing hate speech against migrants and refugees spread through YouTube in Spanish, in: Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21), TEEM'21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 456–460. URL: https://doi.org/10.1145/3486011.3486494. doi:10.1145/3486011.3486494.

[6] K. Yousaf, T. Nawaz, A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos, IEEE Access 10 (2022) 16283–16298. doi:10.1109/ACCESS.2022.3147519.

[7] X. Guo, J. Ma, A. Zubiaga, NUAA-QMUL at SemEval-2020 Task 8: Utilizing BERT and DenseNet for Internet Meme Emotion Analysis, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 901–907. URL: https://aclanthology.org/2020.semeval-1.114. doi:10.18653/v1/2020.semeval-1.114.

[8] K. Maity, P. Jha, S. Saha, P. Bhattacharyya, A Multitask Framework for Sentiment, Emotion and Sarcasm aware Cyberbullying Detection from Multi-modal Code-Mixed Memes, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1739–1749. URL: https://doi.org/10.1145/3477495.3531925. doi:10.1145/3477495.3531925.

[9] D. S. Chauhan, G. V. Singh, N. Majumder, A. Zadeh, A. Ekbal, P. Bhattacharyya, L.-p. Morency, S. Poria, M2H2: A Multimodal Multiparty Hindi Dataset For Humor Recognition in Conversations, in: Proceedings of the 2021 International Conference on Multimodal Interaction, ICMI '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 773–777. URL: https://doi.org/10.1145/3462244.3479959. doi:10.1145/3462244.3479959.

[10] G. K. Shahi, AMUSED: An Annotation Framework of Multi-modal Social Media Data, 2021. URL: http://arxiv.org/abs/2010.00502, arXiv:2010.00502 [cs].

[11] L. Christ, S. Amiriparian, A. Kathan, N. Müller, A. König, B. W. Schuller, Multimodal Prediction of Spontaneous Humour: A Novel Dataset and First Results, 2022. URL: http://arxiv.org/abs/2209.14272, arXiv:2209.14272 [cs, eess].

[12] V. Gupta, T. Mittal, P. Mathur, V. Mishra, M. Maheshwari, A. Bera, D. Mukherjee, D. Manocha, 3MASSIV: Multilingual, Multimodal and Multi-Aspect dataset of Social Media Short Videos, 2022. URL: http://arxiv.org/abs/2203.14456, arXiv:2203.14456 [cs].

[13] S. N. Khan, M. Leekha, J. Shukla, R. R. Shah, Vyaktitv: A Multimodal Peer-to-Peer Hindi Conversations based Dataset for Personality Assessment, 2020. URL: http://arxiv.org/abs/2008.13769, arXiv:2008.13769 [cs].

[14] M. Choubey, Citizen journalism raises hope amid corona virus threats in india, Jamshedpur Res Rev ii (xxxxxi) (2020) 43–49.

[15] M. L. McHugh, Interrater reliability: the kappa statistic, Biochemia medica 22 (2012) 276–282.

[16] S. Zannettou, S. Chatzis, K. Papadamou, M. Sirivianos, The good, the bad and the bait: Detecting and characterizing clickbait on youtube, in: 2018 IEEE Security and Privacy Workshops (SPW), IEEE, 2018, pp. 63–69.

[17] L. Shang, D. Y. Zhang, M. Wang, S. Lai, D. Wang, Towards reliable online clickbait video detection: A content-agnostic approach, Knowledge-Based Systems 182 (2019) 104851.

[18] R. Gothankar, F. D. Troia, M. Stamp, Clickbait detection for youtube videos, in: Artificial Intelligence for Cybersecurity, Springer, 2022, pp. 261–284.

[19] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022. URL: http://arxiv.org/abs/2203.05794, arXiv:2203.05794 [cs].

[20] D. Angelov, Top2Vec: Distributed Representations of Topics, 2020. URL: http://arxiv.org/abs/2008.09470. doi:10.48550/arXiv.2008.09470, arXiv:2008.09470 [cs, stat].

[21] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[22] R. Egger, J. Yu, A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, Frontiers in Sociology 7 (2022) 886498. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/. doi:10.3389/fsoc.2022.886498.

[23] S. Beliga, Keyword extraction: a review of methods and approaches, University of Rijeka, Department of Informatics, Rijeka 1 (2014).

[24] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, A. Jatowt, YAKE! Collection-Independent Automatic Keyword Extractor, in: G. Pasi, B. Piwowarski, L. Azzopardi, A. Hanbury (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 806–810. doi:10.1007/978-3-319-76941-7_80.

[25] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, YAKE! Keyword extraction from single documents using multiple local features, Information Sciences 509 (2020) 257–289. URL: https://www.sciencedirect.com/science/article/pii/S0020025519308588. doi:10.1016/j.ins.2019.09.013.

[26] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic Keyword Extraction from Individual Documents, in: M. W. Berry, J. Kogan (Eds.), Text Mining, John Wiley & Sons, Ltd, Chich-

ester, UK, 2010, pp. 1–20. URL: https://onlinelibrary.wiley.com/doi/10.1002/9780470689646.ch1. doi:10.1002/9780470689646.ch1.

[27] M. Grootendorst, Keybert: Minimal keyword extraction with bert., 2020. URL: https://doi.org/10.5281/zenodo.4461265. doi:10.5281/zenodo.4461265.

[28] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, P. Talukdar, MuRIL: Multilingual Representations for Indian Languages, 2021. URL: http://arxiv.org/abs/2103.10730. doi:10.48550/arXiv.2103.10730, arXiv:2103.10730 [cs].

[29] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL: http://arxiv.org/abs/1810.04805. doi:10.48550/arXiv.1810.04805, arXiv:1810.04805 [cs].

[30] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, 2022. URL: http://arxiv.org/abs/2201.03545. doi:10.48550/arXiv.2201.03545, arXiv:2201.03545 [cs].