Predicting YouTube Hate using Machine Learning

TEAM: (B5) Amish Gautam, Alvaro Tapia Huaman, Areej Soleja

1. BACKGROUND

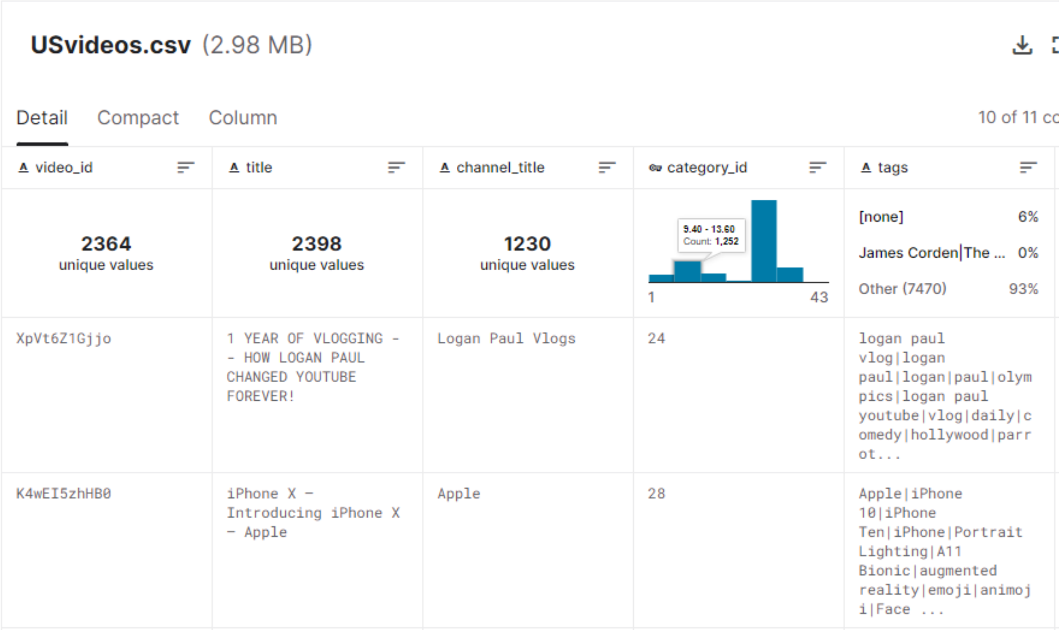
YouTube has been an online video sharing platform for over 15 years now. Around 3.7 million new videos are uploaded each day regarding various topics, issues and genres. However, the platform's open nature has given rise to a growing concern among creators - hate speech and cyberbullying. To alleviate this issue, we have developed a model that predicts whether a YouTube video is likely to receive hateful comments. By analyzing various factors, our model can help creators take proactive measures to protect their mental health and well-being, while also helping YouTube improve its moderation efforts and create a safer and more positive environment for all users.

2. OBJECTIVE

We aim to use the features already available to us in the dataset, along with some extracted features to identify and develop the appropriate machine learning model that helps us predict incoming hate on YouTube with the Highest Accuracy. We are going to be using Data Integration techniques, Sentiment Analysis, TF-IDF Scoring and Machine Learning Modeling.

3. DATASET

We used a dataset of YouTube comments we found through Kaggle. This Dataset contained features like Video ID, Channel Name, Date, Video Category, Video Title, ‘Tags’, Likes, Dislikes, Comment Likes and Comment Replies. We had over 2000 unique videos spread over 15 categories with over 1 million rows of comments. The video dataset and comment datasets very initially separated but we used data integration methods to combine the two.

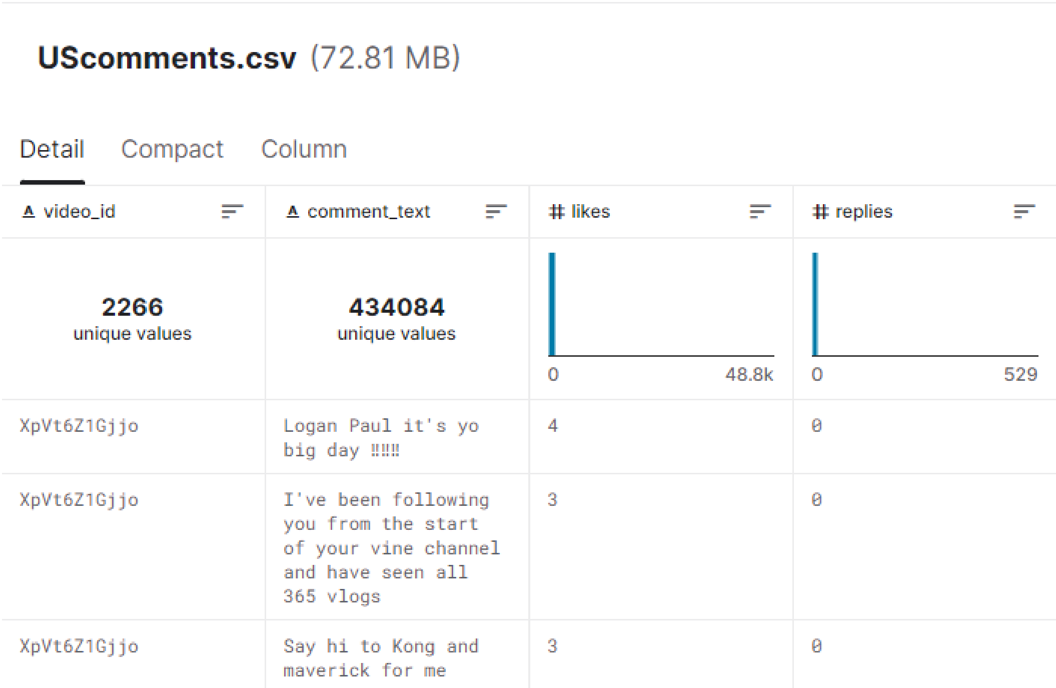


USvideos.csv (2.98 MB)

Detail Compact Column — 10 of 11 cc

video_id	title	channel_title	category_id	tags
2364 unique values	2398 unique values	1230 unique values	9.40 - 13.60 Count: 1,252; 1 … 43	[none] 6%; James Corden|The ... 0%; Other (7470) 93%
XpVt6Z1Gjjo	1 YEAR OF VLOGGING -- HOW LOGAN PAUL CHANGED YOUTUBE FOREVER!	Logan Paul Vlogs	24	logan paul vlog|logan paul|logan|paul|olympics|logan paul youtube|vlog|daily|comedy|hollywood|parrot...
K4wEI5zhHB0	iPhone X — Introducing iPhone X — Apple	Apple	28	Apple|iPhone 10|iPhone Ten|iPhone|Portrait Lighting|A11 Bionic|augmented reality|emoji|animoji|Face ...

Figure 1: US videos dataset



UScomments.csv (72.81 MB)

Detail Compact Column

video_id	comment_text	likes	replies
2266 unique values	434084 unique values	0 … 48.8k	0 … 529
XpVt6Z1Gjjo	Logan Paul it's yo big day ‼️‼️	4	0
XpVt6Z1Gjjo	I've been following you from the start of your vine channel and have seen all 365 vlogs	3	0
XpVt6Z1Gjjo	Say hi to Kong and maverick for me	3	0

Figure 2: US comments dataset

4. DATA PRE-PROCESSING

We first had to combine the two datasets in order to create a master table that had all the features we were supplied with through Kaggle. We then cleaned the dataset by dropping duplicate rows and dropping the features we did not think we needed. We also created many sub-datasets that we would be putting through future methods for feature extraction for better processing time.

	video_id	title	channel_title	category_id	tags	views
0	XpVt6Z1Gjjo	1 YEAR OF VLOGGING -- HOW LOGAN PAUL CHANGED Y...	Logan Paul Vlogs	24	logan paul vlog|logan paul|logan|paul|olympics...	4394029
801	cLdxuaxaQwc	My Response	PewDiePie	22	[none]	5845909
1600	WYYvHb03Eog	Apple iPhone X first look	The Verge	28	apple iphone x hands on|Apple iPhone X|iPhone ...	2642103
2400	sjlHnJvXdQs	iPhone X (parody)	jacksfilms	23	jacksfilms|parody|parodies|iphone|iphone x|iph...	1168130
3200	cMKX2tE5Luk	The Disaster Artist | Official Trailer HD | A24	A24	1	a24|a24 films|a24 trailers|independent films|t...	1311445
...	...	...	...	...	...	...
2900192	mv4MRmwXJMM	Kygo - Kids in Love (Audio) ft. The Night Game	KygoOfficialVEVO	10	Dance/House/Techno|Kids in Love|Kygo|Kygo feat...	1249946
2903564	7_GaeAoLMWY	Keyshia Cole Performs Incapable	Wendy Williams	24	Keyshia Cole|wendy williams|the wendy williams...	106467
2924730	S9VIKOuZcds	My Sweet Jax (Tribute to a Cat)	Hot Dad	23	cat|feline|pets|beloved|family member|grief|gr...	25037
2925632	a5Nlg5yyHWo	Pawn Stars: An Original 1978 Superman Costume ...	HISTORY	24	history|history channel|history shows|history ...	400104
2937660	3VSa-oARk-w	Monument Valley 2 - Available on Android & iOS...	ustwo games	20	Monument Valley|Optical Illusions|Gaming|Mobil...	24075

2266 rows × 14 columns

Figure 3: Merged Datasets

5. METHODOLOGY

To get our data prepared to be put through machine learning models, we first had to create valuable features. We used the following methods:

- **Sentiment Analysis**: We used an NLTK Sentiment Analyzer and ran it through our dataset's video comments column. It gave us a score between -1 and 1 for each comment, which we then found the average of per video and created a separate 'average_sentiment' column for. Once we did that, we dropped the comments column, along with all the other rows for each video except the one with the highest views. This ensured we had the YouTube video's latest row with its average sentiment. We also created a 'sentiment_label' feature that assigned binary values to a video which implied if it received hate or not.

comment_sentiment	average_sentiment	sentiment_label
0.0000	0.086394	1
0.5126	-0.020132	0
0.0000	0.070321	1
0.0000	0.115051	1
0.3182	0.179144	1
...	...	...
0.8930	0.253038	1
0.7717	0.313223	1
-0.9062	0.144990	1
0.8591	0.198645	1
0.0000	0.074111	1

Figure 4: Visualization of Sentiments

- **TF-IDF**: We then used the sub-table we created for the 'tags' column which also contained the video_id for that tag and ran a TF-IDF Scoring Model on it so each tag used in a video was assigned a score based on the number of times it occurred in the tags column, and all the scores per tag per video were added up to a final 'tfidf_sum' column. We did this so we could identify which videos used more important or heavy weighted tags.

sentiment_label	tfidf_sum
1	2.937741
1	2.247886
1	2.259244
1	2.909701
1	3.946778

Figure 5: Visualization of TF IDF

- **Correlation matrix**: Implementation of correlation matrix in order to find potential correlations and patterns between the features. Potential findings were that even though there isn't a strong correlation between most features, the sentiments and TF IDF features do have a good correlation compared to other features. ChatGPT explains this good since our datasets were extremely large.

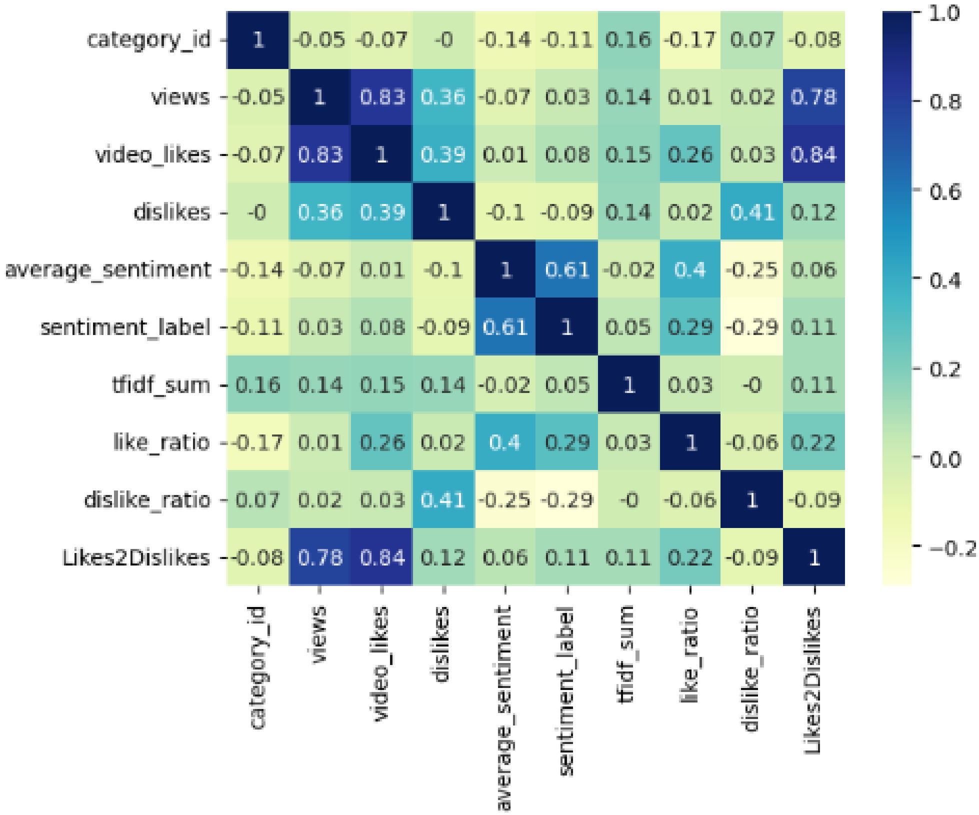


Figure 6: Correlation Matrix

- **Feature Extraction:** We used features that were already provided to us like views, likes and dislikes and calculated ratios for each of those to be used as useful features in our model.

video_id	channel_title	category_id	views	video_likes	dislikes	average_sentiment	like_ratio	dislike_ratio	Likes2Dislikes	sentiment_label	tfidf_sum
XpVt6Z1Gjjo	Logan Paul Vlogs	24	4394029	320053	5931	0.129263	0.072838	0.001350	2.371138e+08	1	2.937741
WYYvHb03Eog	The Verge	28	2642103	24975	4542	0.113658	0.009453	0.001719	1.452808e+07	1	2.247886
sjlHnJvXdQs	jacksfilms	23	1168130	96666	568	0.111827	0.082753	0.000486	1.988001e+08	1	2.259244
cMKX2tE5Luk	A24	1	1311445	34507	544	0.108009	0.026312	0.000415	8.318756e+07	1	2.909701
8wNr-NQImFg	Late Night with Seth Meyers	23	666169	9985	297	0.111827	0.014989	0.000446	2.239629e+07	1	3.946778
...	...	...	...	...	...	...	...	...	...	...	...
fc_oYX7JJ-U	Harper's BAZAAR	26	8544	95	0	0.300629	0.011119	0.000000	9.999900e+04	1	1.990241
dUBcP00TEWI	Lucas	23	186354	11129	276	0.111827	0.059720	0.001481	7.514252e+06	1	1.726343
S9VIKOuZcds	Hot Dad	23	25037	2846	11	0.111827	0.113672	0.000439	6.477755e+06	1	1.728996
a5NIg5yyHWo	HISTORY	24	400104	2432	123	0.129263	0.006078	0.000307	7.910999e+06	1	1.000000
3VSa-oARk-w	ustwo games	20	24075	189	2	0.146266	0.007850	0.000083	2.275088e+06	1	1.401928

Figure 7: Dataset after Feature Selection

- **Category Sentiment Analysis**: Implementation of the comparison between category of a video vs. average sentiment. The purpose of this was to be able to identify what type of videos tend to be the least accepted by the community and also which ones are the most accepted. With this, we are able to identify in the figure above that News and Politics tend to be the most hated videos while Nonprofit videos are better accepted by the people.

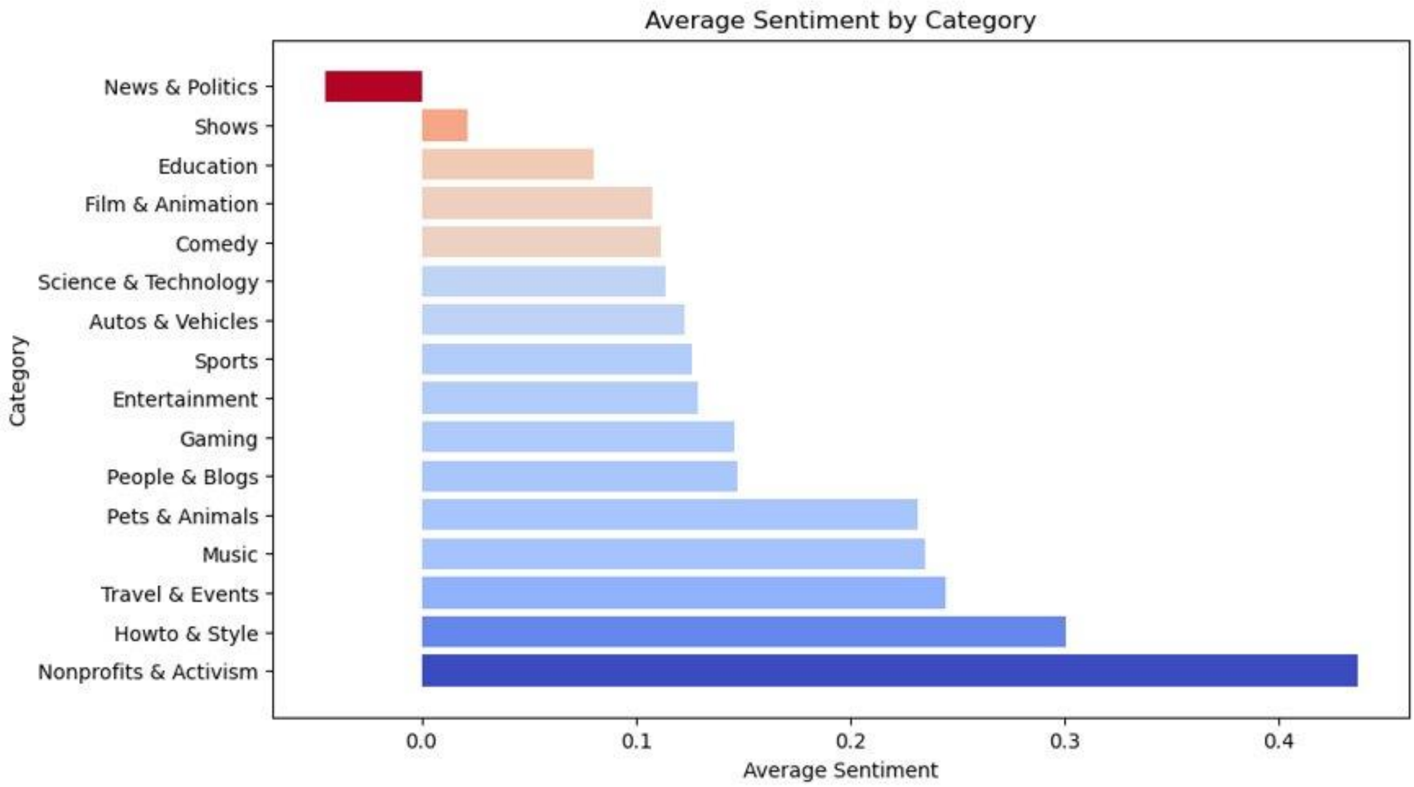


Figure 8: Correlation Matrix

- **Machine Learning:** After successful feature extraction, we explored various machine learning modules like Logistic Regression, Support Vector Machines and Random Forest Classifier. We randomly split our data into testing and training sets with an 80:20 split. After fitting the best features and running the various models, we found that Random Forest Classifier gave us the best accuracy of 0.9671 or 96.7%. In our hyper parameter tuning, we used the number of trees or n_estimators to be 100.

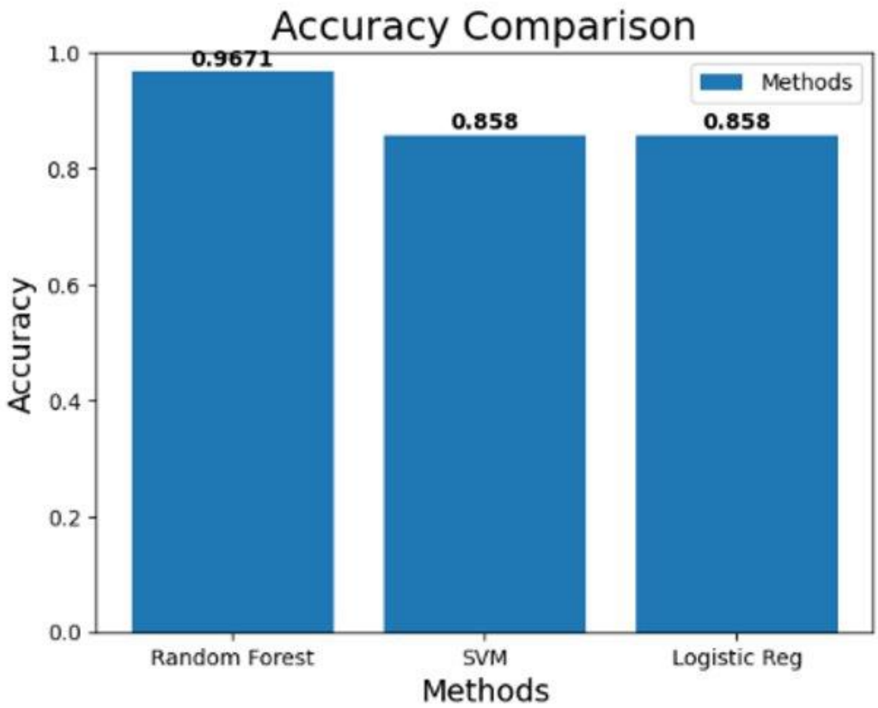


Figure 9: Accuracy comparison between Models

	category_id	views	video_likes	dislikes	average_sentiment	like_ratio	dislike_ratio	tfidf_sum	predicted_sentiment
387	25	59450	398	63	0.027112	0.006695	0.001060	1.000000	1
1426	24	2746174	78698	13980	-0.016123	0.028657	0.005091	3.326209	0
697	10	2781186	370543	1193	0.245484	0.133232	0.000429	2.512246	1
1739	24	262396	2914	94	0.208382	0.011105	0.000358	4.143615	1
735	22	189698	13351	82	0.239241	0.070380	0.000432	2.980937	1
...	...	...	...	...	...	...	...	...	...
123	17	113400	1049	20	0.257753	0.009250	0.000176	1.726604	1
2193	17	822925	11584	713	0.044986	0.014077	0.000866	1.413067	1
136	25	171000	454	85	0.127661	0.002655	0.000497	2.209042	1
2111	26	347621	17286	237	0.232200	0.049727	0.000682	6.253642	1
2243	24	477756	30678	207	0.151470	0.064213	0.000433	2.596606	1

Figure 10: Model Fitted in X_test Dataframe

- **World Cloud:** We additionally included a world cloud to better identify what are the words mostly used for videos, and also what words are the less accepted.

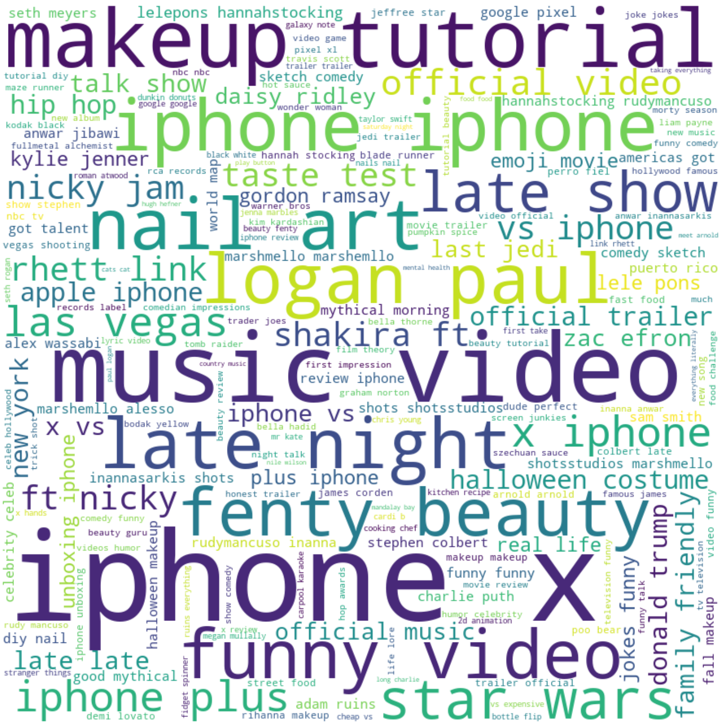


Figure 11: World Cloud most popular words

Figure 12: World Cloud most hated words

- **CHAT-GPT Usage:** Throughout this project, we made great use of CHAT-GPT. It was the tool that gave our project suggestions for the models early on in the project. It also guided us to the methods of performing TF-IDF scoring and Sentiment Analysis. We kept asking it ways to improve at every stage and it was crucial to our project.

6. CONCLUSION

Upon evaluating our model, we were pleased to find that it achieved an impressive accuracy of 96.71%. Although we recognize that our sample size was limited to around 2000 videos, we are confident that this accuracy is indicative of the model's potential to perform even better with a larger dataset.

Our analysis also revealed that the News and Politics, Shows, and Education categories were the top three categories receiving the most hate speech. This information provides valuable insights for identifying the areas that require more attention to prevent online hate. We aim to use these findings to guide our efforts in creating a safer and more positive environment on YouTube.

REFERENCES

"Trending YouTube Video Statistics and Comments." Kaggle, 24 Nov. 2017, https://www.kaggle.com/datasets/datasnaek/youtube?select=UScomments.csv

"Trending YouTube Video Statistics and Comments." Kaggle, 24 Nov. 2017, https://www.kaggle.com/datasets/datasnaek/youtube?select=USvideos.csv

"List of YouTube Video Category IDs." MixedAnalytics, 31 Dec. 2022, https://mixedanalytics.com/blog/list-of-youtube-video-category-ids/