Research Questions: *Does YouTube have any impact on the research community? If so, can we measure this impact using video citations? What other YouTube features contribute to the impact of the research articles on social media platforms? Do certain categories of videos generate negative or positive emotions? Which category of videos are people generally inclined to like? Are the count of views correlated to the video citations? Are the views or likes count correlated to the subscriber counts? Given a research paper, will there be a YouTube video that links to it?*

Introduction:

Social computing platforms have seen a tremendous growth in the past decade. Various platforms such as Twitter, Facebook, Mendeley and YouTube have shifted the role of how information is traversed through the internet. This shift has caused millions of posts, tweets and video uploads on different platforms, instigating a huge impact in the field of scholarly research. The area of videos citing research literature on the YouTube platform is new and has potential for impactful research. Our work includes data collection from different databases and machine learning models would be built on them to predict the different features collected from YouTube. The dataset would come from Altmetric.com and YouTube, we would collect various features and analyze the dataset using different statistics and visualizations leading to a machine learning model which would be able to predict the impact of videos in the field of scholarly research.

Susarla et al. [1] stated that social media has become the main platform for real time updates on various news throughout the world. YouTube has been the primary source for video sharing since its initiation and billions of people use it. Snelson et al. [2] stated that YouTube is the highest online video sharing platform and is also the most under-researched platform in comparison to other social media platforms. Chau [3] stated that YouTube provides a platform for the youth to become media creators and reach a global audience, delivering a place to collaborate, learn and circulate. Madathil et al. [4] conducted a review on 18 articles relating to YouTube and Healthcare finding that health information is increasingly being conveyed through YouTube and allows users from all around the globe to view, communicate and upload health information. Although the trustworthiness of some of this health information on the social media platform is vague, there is huge potential for YouTube to become an efficient information source for healthcare.

Related Work:

Snelson et al. [2] studied YouTube using the Delphi method with an expert panel to identify and prioritize seven future research areas of YouTube concluding that human behavior and social interaction on YouTube is the highest research priority category followed by education and social impact. Susarla et al. [1] indicated that accounts that are old on YouTube and have more videos posted have higher chances of causing an impact. Jaffar [5] led a study on 91 medical students to assess the effectiveness of YouTube videos towards a given anatomy problem and found that 98% of students used YouTube extensively and 92% of students agreed on YouTube helping them learn anatomy. YouTube can be used as an effective and efficient learning tool given that the videos are trustworthy, evaluated and aimed towards the course description.

Behavioral research has been conducted on YouTube to understand the social and non-social impact. Gigletto et al. [6] proposed a literature review using Facebook, Twitter and YouTube data for behavioral research of the use of social media. They also discussed about the limitations of social media data and how to evaluate the usage of information publicly available. Khan and Vong [7] analyzed top viral videos by building an empirical model to understand how videos achieve virality and the relationship between different aspects such as social and non-social capital. They found that offline social capital and network dynamics contribute more towards the virality along with the view count. Moor [8] conducted a survey to understand the act of flaming, demonstrating hostility by insulting or swearing, on YouTube concluding that flaming is a result of several aspects such as disagreement or response to flaming. June et al. [9] led an action research on 50 students to assess the critical thinking skills and interactive activities of the students while using videos on YouTube discovering that it has a huge potential to serve as a tool for students learning since it enhanced the learning experience and improved critical thinking skills of the students.

Approach: The database built for this study comes from two data sources Altmetric.com and YouTube. The final dataset includes a combination from these data sources mounting up to around 20 features consisting of social media mentions of the research articles and meta data of the YouTube videos citing the literature. Using this dataset, visualization techniques and statistical models would be applied for analysis to come up with a machine learning model predicting the video citation of scholarly research using different social media mentions.

Expected Result: A predictive model to predict the video citations and metadata of videos citing scholarly research on YouTube

Data Extraction:

A dataset of 1,000,000 sample research papers were collected from Altmetric Explorer which included metadata of the papers and the mentions of research outputs on various social media networks. There were a total of 121,928 research papers cited in videos on YouTube from the sample collected, a separate database was created which contained altmetric IDs, metadata of the research paper, post count from various social media platforms such as twitter, facebook, news, policy, reddit and youtube. The YouTube video ids of those research papers were collected which cited any scholarly work on the YouTube platform. There were 89,420 links collected among which 88,928 links were unique and among these 87,009 links were available. Using the YouTube video IDs collected, another dataset was created containing the YouTube video ID, title, views, likes, dislikes, category, description and number of counts (metadata of the video). Finally, both of these datasets were combined to form a full complete dataset consisting of all research papers cited on YouTube with the links to the videos and the metadata of the videos. This dataset contained a final count of 121,928 research papers with the metadata of the papers and the information of all the YouTube videos citing the scholarly work.

A complete description of the features in all the three datasets :

Dataset 1 - YouTube :

Links and metadata of all the videos citing a scholarly research.

Features - link, title, views, likes, dislikes, channelname, subno, pubdate, description, category, commentcount

Link - Youtube unique link or ID of the video
Title - Title of the YouTube video
Views - The total views of the YouTube video
Likes - The total views of the YouTube video
Dislikes - The total views of the YouTube video
Channelname- Name of the channel that posted the video
Subno - The total subscriber count of the channel
Pubdate - Date of the video when it was published
Description - Description of the video as provided by the channel
Category - Category of the video as mentioned
CommentCount - The total comments posted on the YouTube video

Dataset 2 - Altmetrics :

Altmetrics of the scholarly research cited in a YouTube video

Features - altmetricID, alt_title, abstract, subjects, Facebook_citations, Google+_citations, Reddit_citations, Twitter_citations, YouTube_citations, Dimension_citations, Wikipedia_citations

AltmetricID - Altmetric ID of the scholarly research
Alt_title - Title of the scholarly research
Abstract - Abstract of the scholarly research
Subjects - Subject of the scholarly research
Facebook_citations - Total citations of the scholarly research on Facebook
Google+_citations - Total citations of the scholarly research on Google+
Reddit_citations - Total citations of the scholarly research on Reddit
Twitter_citations - Total citations of the scholarly research on Twitter
YouTube_citations - Total citations of the scholarly research on YouTube
Dimension_citations - Total citations of the scholarly research on Dimensions
Wikipedia_citations - Total citations of the scholarly research on Wikipedia

Dataset 3 - Scholarly research

A combination of Dataset 1 and Dataset 2 with each scholarly work containing metadata of all the YouTube videos citing the scholarly work and altmetrics for the scholarly research.

Features - altmetricID, alt_title, abstract, subjects, Facebook_citations, Google+_citations, Reddit_citations, Twitter_citations, YouTube_citations, Dimension_citations,

Wikipedia_citations [link], [title], [views], [likes], [dislikes], [subname], [subno], [pubdate], [description], [category], [commentcount]

[--] describes a feature with multiple values.

The features of videos in this dataset have multiple values as each scholarly work had video citations ranging from a count of 1 to 789 videos. The description of the features is similar to the ones stated above.

Dataset 4 - YouTube Videos

A combination of Dataset 1 and Dataset 2 with each YouTube video containing metadata of all the scholarly research cited by the video and altmetrics for the scholarly research.

Features - link, title, views, likes, dislikes, subname, subno, pubdate, description, category, commentcount, [altmetricID], [alt_title], [abstract], [subjects], [Facebook_citations], [Google+_citations], [Reddit_citations], [Twitter_citations], [YouTube_citations], [Dimension_citations], [Wikipedia_citations]

[--] describes a feature with multiple values.

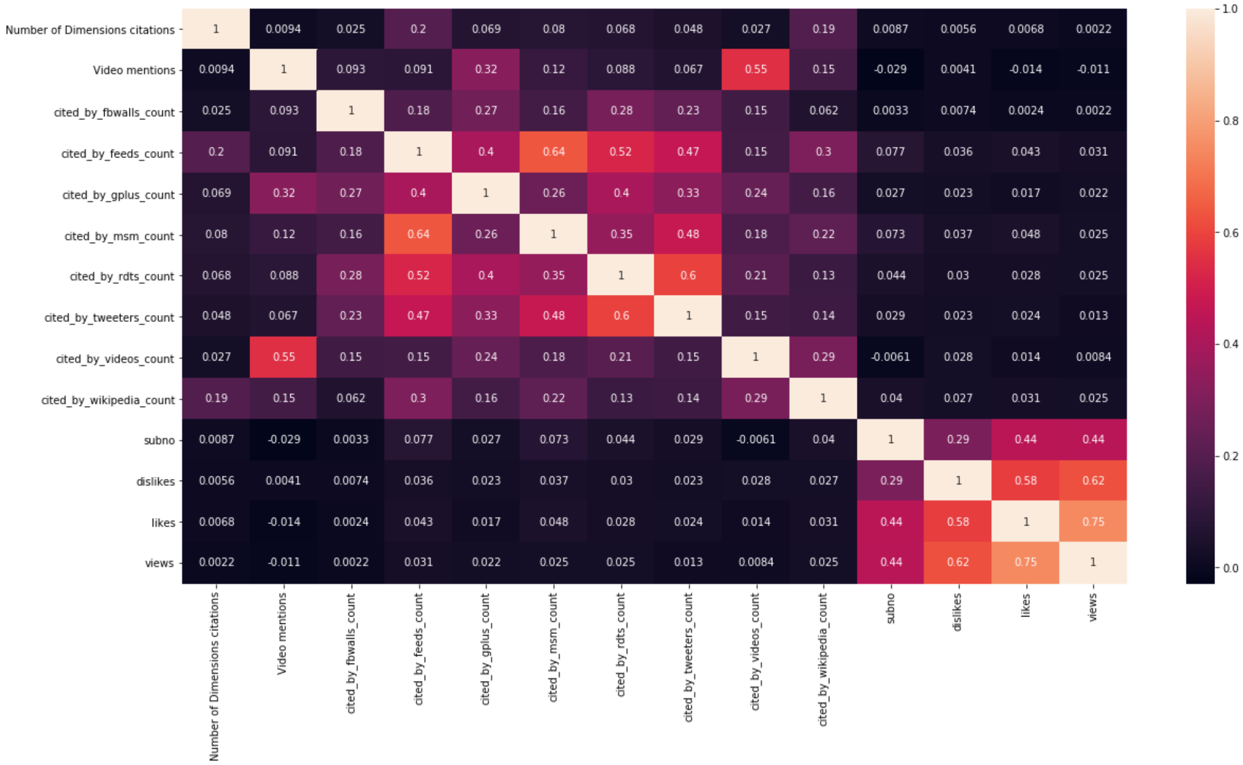
The features of altmetrics in this dataset have multiple values as each YouTube video had citations of various scholarly research. The description of the features is similar to the ones stated above.

Data Preprocessing:

Each of the three datasets used in the project underwent preprocessing to ensure removal of abnormalities and null values in the dataset for optimized performances. In **Dataset 1**, containing links and the metadata of videos, we removed those links with no views data. There was a total of 10564 videos with 0 or NULL values for views. The dataset was updated to 78,898 videos, among which 2370 had ‘No views’ values in likes and dislikes feature. These values were converted to 0 views. Most of the features were converted from string to integer values removing unnecessary data.

Basic statistics:

Correlation matrix for all the features in the data set



	Number of Dimensions citations	Video mentions	cited_by_fbwalls_count	cited_by_feeds_count	cited_by_gplus_count	cited_by_msm_count	cited_by_rdts_count	cited_by_tweeters_count	cited_by_videos_count	cited_by_wikipedia_count	subno	dislikes	likes	views
Number of Dimensions citations	1	0.0094	0.025	0.2	0.069	0.08	0.068	0.048	0.027	0.19	0.0087	0.0056	0.0068	0.0022
Video mentions	0.0094	1	0.093	0.091	0.32	0.12	0.088	0.067	0.55	0.15	-0.029	0.0041	-0.014	-0.011
cited_by_fbwalls_count	0.025	0.093	1	0.18	0.27	0.16	0.28	0.23	0.15	0.062	0.0033	0.0074	0.0024	0.0022
cited_by_feeds_count	0.2	0.091	0.18	1	0.4	0.64	0.52	0.47	0.15	0.3	0.077	0.036	0.043	0.031
cited_by_gplus_count	0.069	0.32	0.27	0.4	1	0.26	0.4	0.33	0.24	0.16	0.027	0.023	0.017	0.022
cited_by_msm_count	0.08	0.12	0.16	0.64	0.26	1	0.35	0.48	0.18	0.22	0.073	0.037	0.048	0.025
cited_by_rdts_count	0.068	0.088	0.28	0.52	0.4	0.35	1	0.6	0.21	0.13	0.044	0.03	0.028	0.025
cited_by_tweeters_count	0.048	0.067	0.23	0.47	0.33	0.48	0.6	1	0.15	0.14	0.029	0.023	0.024	0.013
cited_by_videos_count	0.027	0.55	0.15	0.15	0.24	0.18	0.21	0.15	1	0.29	-0.0061	0.028	0.014	0.0084
cited_by_wikipedia_count	0.19	0.15	0.062	0.3	0.16	0.22	0.13	0.14	0.29	1	0.04	0.027	0.031	0.025
subno	0.0087	-0.029	0.0033	0.077	0.027	0.073	0.044	0.029	-0.0061	0.04	1	0.29	0.44	0.44
dislikes	0.0056	0.0041	0.0074	0.036	0.023	0.037	0.03	0.023	0.028	0.027	0.29	1	0.58	0.62
likes	0.0068	-0.014	0.0024	0.043	0.017	0.048	0.028	0.024	0.014	0.031	0.44	0.58	1	0.75
views	0.0022	-0.011	0.0022	0.031	0.022	0.025	0.025	0.013	0.0084	0.025	0.44	0.62	0.75	1

Research questions:

Does the YouTube have any impact on the research community? If so, can we measure this impact using video citations or mentions? What are the YouTube features contributing to the impact of the research articles on social media platforms?

Models:

Predicting the dimension citations of a research paper.

The dataset considered is of 121K research papers which has YouTube video mentions. Features considered for this model are number of video mentions, Facebook posts, Googleplus, Twitter, Wikipedia, Reddit, News, Blog from the altmetric data and average number of views, likes, dislikes, comments of videos which has research paper mentions.

Target Variable:

Papers which has dimension citations more than the median of the entire dataset are considered as Class Label ‘1’ and the remaining are ‘0’. With the mentioned features and target, the dataset is divided into train and test set of 75% and 25%. Following models were trained and the test results are

Model	Precision	Recall	Accuracy
Random Forest	0.6	0.6	0.60
Logistic Regression	0.61	0.58	0.58

Predicting views per day for a YouTube video citing a research paper.

The dataset considered is of 88K videos which are citing a research paper. Features considered for this model are number of Dimension citations, video mentions of a paper, Facebook posts, Googleplus posts, Twitter posts, wikipedia mentions, reddit posts, number of subscribers of a channel posting the video.If a video has more than one research paper citations, the average of above features was taken as input.

Target:

From number of views of a video and number of days from the published data, number of views per day is calculated. Videos which has got number of views per day greater than the median of entire data set is given class label of *ONE* and which are less than the median is considered class *ZERO.*

With the mentioned features and target, the dataset is divided into train and test set of 75% and 25%. Following models were trained and the test results are

Model	Precision	Recall	F1 score	Accuracy
Random Forest	0.81	0.81	0.81	0.80
Logistic Regression	0.78	0.73	0.71	0.73
Decision Tree	0.78	0.77	0.76	0.77

Predicting the category of YouTube videos:

From the dataset of 88K videos, the videos belong to 1360 categories. Out of the entire dataset 87% of videos are in Education, People & Blogs, Entertainment, How to & Style, Science & Technology, Sports categories. The dataset is filtered choosing the mentioned categories, the size of the dataset is now 77K videos. Model was built to predict the category of the paper.

Input features for the model are average number of dimension citations, video mentions, Facebook posts, Googleplus, Reddits, Twitter, Wikipedia of all the research papers which were cited by a YouTube video and number of subscribers to the channel posted to the video and views, likes, dislikes of a video.

Target:

Category of the video mentioned.

Models trained and tested on the data.

Model	Precision	Recall	F1 Score	Accuracy
Random Forest	0.64	0.65	0.64	0.65
Decision Tree	0.39	0.43	0.37	0.43
Logistic Regression	0.36	0.42	0.31	0.42

Predicting the subject of research paper using altmetric features and YouTube features.

Out of the 121k papers 81k papers have subjects mentioned to it. All the research papers were belonged to total of 30 subjects. Some of the subjects were grouped together to reduce the number of subjects from 30 to 7. Those seven subjects are considered as target variables. Each paper may belong to one or more subjects, making this as multi-class and multi-label classification problem.

Input features for this model are number dimension citations, video mentions, Facebook posts, Googleplus, Twitter, Wikipedia, Reddit, News and Blog mentions from the altmetric data and average views, likes, dislikes, number of comments, subscriber number of channels posted.

Models trained and tested on:

Model	F1 Score	Hamming loss
Random Forest	0.72	0.144
Decision Tree	0.73	0.135
Logistic Regression	0.64	0.18

The same analysis was done by adding the text analysis on the abstracts of the research papers. Form 81k papers only 51k papers have abstracts in the dataset. The texts of the abstracts are cleaned and processed using NLP techniques such as stemming and lemmatization. Using LDA (Latent Dirichlet Allocation) all the abstracts were assigned a topic against the weights of the words. Each research paper was given a score against topics (a total of 10 topics were chosen for topic modelling). These scores were taken as input features for the above-mentioned models. (For example, a research paper has a score of 0.87 for topic 2 and 0.13 topic 10). We could see a significant improvement in the results for the models trained and tested.

Models trained and tested after topic modeling on abstracts in research papers:

Model	Precision (micro average)	Recall (micro average)	F1 Score	Hamming loss	Accuracy
Random Forest	0.83	0.85	0.84	0.082	0.67
Logistic Regression	0.83	0.81	0.85	0.079	0.63

Similarity scores between video transcripts and abstracts of research paper

A sample of 5746 video transcripts were extracted from YouTube. These videos were belonged to total of 14729 research papers. Cosine similarity of abstract and video transcripts were calculated. Out of 14729 17% papers have similarity scores greater than 0.5 and 65% papers have similarity scores between 0.5 and 0.25 and the rest have scores less than 0.25.

Conclusion

Analysis of YouTube impact on the research community has been showcased the different machine learning models built which can predict important factors of the research community. Future work includes but is not limited to including the transcript of the videos as a much more detailed feature to further explore the YouTube analysis on research community. And also to find the relevance of content in videos to the cited research paper.

References

[1] A. Susarla, J.-H. Oh, and Y. Tan, "Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube," *Information Systems Research*, vol. 23, no. 1, pp. 23–41, Mar. 2012.

[2] C. Snelson, K. Rice, and C. Wyzard, "Research priorities for YouTube and video-sharing technologies: A Delphi study," *Br. J. Educ. Technol.*, vol. 43, no. 1, pp. 119–129, 2012.

[3] C. Chau, "YouTube as a participatory culture," *New Dir. Youth Dev.*, vol. 2010, no. 128, pp. 65–74, Winter 2010.

[4] K. C. Madathil, A. J. Rivera-Rodriguez, J. S. Greenstein, and A. K. Gramopadhye, "Healthcare information on YouTube: A systematic review," *Health Informatics J.*, vol. 21, no. 3, pp. 173–194, Sep. 2015.

[5] A. A. Jaffar, "YouTube: An emerging tool in anatomy education," *Anat. Sci. Educ.*, vol. 5, no. 3, pp. 158–164, May 2012.

[6] F. Giglietto, L. Rossi, and D. Bennato, "The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source," *J. Technol. Hum. Serv.*, vol. 30, no. 3–4, pp. 145–159, Jul. 2012.

[7] G. Feroz Khan and S. Vong, "Virality over YouTube: an empirical analysis," *Internet research*, vol. 24, no. 5, pp. 629–647, 2014.

[8] P. J. Moor, A. Heuvelman, and R. Verleur, "Flaming on YouTube," *Comput. Human Behav.*, vol. 26, no. 6, pp. 1536–1546, Nov. 2010.

[9] S. June, A. Yaacob, and Y. K. Kheng, "Assessing the use of YouTube videos and interactive activities as a critical thinking stimulator for tertiary students: An action research," *International Education Studies*, vol. 7, no. 8, pp. 56–67, 2014.