International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 2, pp 123-127, February-2024

https://doi.org/10.47001/IRJIET/2024.802017

Data Mining Method for Video Subscribers and Analysis of Youtube

[1]Nadhiya S, [2]S. Dinesh Kumar, [3]Dr. N. Sudha, [4]Dr. G. Chitra Ganapathy

[1]PG Student, Department of Computer Science and Engineering, CMS College of Engineering, Tamilnadu, India
[2]Professor, Department of Computer Science and Engineering, CMS College of Engineering, Tamilnadu, India
[3]Principal, CMS College of Engineering, Tamilnadu, India
[4]HoD, Department of Computer Science and Engineering, CMS College of Engineering, Tamilnadu, India

***Abstract* - With a huge collection of videos attracting billions of views from users each month, the resulting data generated by YouTube is enormous. Data like view count, like count, dislike count, user comments, etc are all valuable data that can be extracted and analyzed to uncover insights about user preferences and sentiment towards a particular video or a particular cause e.g. the Ice Bucket Challenge, sometimes called the ALS Ice Bucket Challenge that went viral on YouTube few years ago. It also presents valuable information to marketers in their decision-making process of promoting a particular product or service. A fun example would be a movie studio, having uploaded a new movie trailer on their YouTube channel and would like to know about viewers' response towards the upcoming movie. Statistics on the movie trailer such as view count, like count and user comments can help marketers to gauge the market response to the movie and allocate their marketing budget accordingly. This is the primary motivation behind this project. In this project, I have extracted and analyzed some interesting statistics about popular superhero movies from both Disney/Marvel and Warner Bros/DC such as Infinity War, Justice League, Black Panther, Wonder Woman and Aquaman.**

Keywords: YouTube; Trending Video; Statistics; Causality; Video Analysis, Data Mining, Sentimental Analysis.

I. INTRODUCTION

Videos contain a vast amount of information, which, if channeled properly can provide breakthrough in various research fields. Multimedia content retrieval is an important research field that aims in content based information indexing and retrieval, automatic annotation and structuring of video frames. Large amount of information is embedded in Natural Scene, which often requires automatic extraction and processing; Artificial Text can be termed as one of the important Multimedia contents. In this Paper, it is attempted to accurately separate text content from multimedia objects by precisely localizing the text and extracting. Text detection in videos is an important step to achieve Multimedia content retrieval which plays an important part in fully understanding of the video clip. Object retrieval follows the general procedure involving detection, localization, tracking, extraction and enhancement of the text from a given image. The detection step roughly classifies text & non-text regions, the localization step determine accurate boundaries of the text string, the extraction step filters out background pixels in the text string.

Images and videos on webs and in databases are increasing. It is a pressing task to develop effective methods to manage and retrieve these multimedia resources by their content. Text, which carries high-level semantic information, is a kind of important object that is useful for this task. The acquisition of a video is generally done using a set of physical captors, Web camera or a Digital Camera, in this Paper. The next step of separating image frames can be accurately modeled as a sampling of the continuous image using a discrete partitioning of the continuous plane. The Image frame thus acquired is the “Digital Image” and the basic procedures of Digital Image Processing are applied. In the case of video-clips; the number of frames containing text is much smaller than the number of frames without text. Binary images considered for analysis generally result from the digitization of frames obtained from Video clip. Similarly, such binary images can be created by thresholding the grey-level at each pixel in grey scale images. This processing represents the most basic operation in the class of segmentation processes.

Data Encryption exchanges between nodes in the wide network are widely used to ensure data security.

In this project, I have used YouTube Data API v3 to extract data about videos and retrieve their statistics such as number of views, likes, dislikes, comments, etc. Then I converted these statistics into pandas Data Frame for further analysis. Analysis performed included ranking the most popular videos by view count and like count, and analyzing view count by day of the week. In the last part of this project, I have also extracted the comments from a YouTube video and

ISSN (online): 2581-3048
Volume 8, Issue 2, pp 123-127, February-2024
https://doi.org/10.47001/IRJIET/2024.802017

performed word cloud visualization on the most popular words, and followed by sentiment analysis using NLTK. The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

To use the YouTube Data API, you will need a API key which you can obtain in the Google APIs Console. You will need to create a new project, and enable the YouTube Data API for the project. To simplify things, I have used Google API python client which is a python client to interact with the Google APIs in an easier way. In order to use the API, you have to build a resource object for that API. Alternatively, you can also manually call the YouTube Data API to retrieve the data, by setting the query parameters in the API endpoint to perform search on query term that you are interested.

YouTube is the world's most renowned electronic video site, with customers watching 4 billion hours of video consistently, and moving 72 hours of video reliably (YouTube, 2013). YouTube began in February 2005 and was set up by Chad Hurley, Steve Chen, and Jawed Karim who named it "YouTube.com". Through the YouTube stage, people started to make a video-sharing site on which customers could move, deal, and view accounts. From here on out, YouTube has obtained a horde of individuals of billions of customers including educators and scientists. Data Analysis and Mining are becoming fundamental to receive the significant examples in return. Regardless, an enormous piece of the Data made is generally in Huge Size and comes in unstructured course of action. Enormous Data can't be examined by ordinary informational collection structures and cycles. To decide this issue, numerous new gadgets that execute Parallel Processing are being sent in these affiliations.

We utilized informational collection from YouTube using YouTube api of a particular day and performed Data Analysis on the data to comprehension into latest examples and customer responsibility in YouTube concerning Classes. Data Investigation and Visualization was done using Google Collaboratory. Examination of coordinated data has seen colossal achievement previously. In any case, assessment of huge extension unstructured data as video configuration remains a troublesome area. YouTube, a Google association, has north of a billion customers and produces billions of points of view. Since YouTube data is getting made in an especially colossal aggregate and with a correspondingly extraordinary speed, there is a gigantic interest to store, process and circumspectly focus on this tremendous proportion of data to make it usable.

II. PROBLEM FORMULATION

As the access to computerized correspondence turns out to be progressively basic around the globe, it ends up less demanding to trade thoughts between societies, crossing once-restrictive geographic separations and national limits. Online video is especially interesting as a potential vector for social correspondence. As a visual resource, video offers the possibility to cross semantic and education hindrances. Video resources are additionally ready to catch social encounters, for example, move and music, which are hard to pass on through content-based media. Information mining is an expansive region that incorporates strategies from a few fields including AI, insights, design acknowledgment, man-made consciousness, and database frameworks, for the investigation of substantial volumes of information. There have been countless mining calculations attached in these fields to perform distinctive information examination tasks. Google give most intriguing thing and significant thing which is valued by each one whose most youthful to most established one in the time of present-day era. Google gives that thing is. YouTube is where sound and video has been seen by the clients. Loads of audio and recordings materials are on YouTube. Every client need to watch their fascinating sound and recordings material they visit. YouTube is where there is no bar of age or sexual orientation. Each one inquiries their very own decision.

Information Mining or “the proficient disclosure of important, on-evident data from a huge accumulation of data” has an objective to find learning out of information and present it in a structure that is effectively fathomable to people. Learning recognition in databases is an exact procedure comprising of various unmistakable advances. Information mining is the establishment step, which results in the revelation of obscure however accommodating learning from tremendous databases. A formal meaning of knowledge revelation in databases is given as pursues: “Information mining, or learning disclosure, is the PC helped procedure of burrowing through and breaking down tremendous arrangements of information and after that extricating the significance of the information. Information mining devices anticipate practices and future patterns, enabling organizations to make proactive, learning driven choices. Data mining ability gives a buyer inclining way to deal with new and obscure examples in the information.

The uncovered learning can be utilized by the human services directors to advance the prevalence of administration. Information mining is a wide zone that coordinates strategies from a few fields including AI, measurements, design acknowledgment, man-made brainpower, and database frameworks, for the examination of substantial volumes of information. It is normally utilized in a wide scope of profiling rehearses, for example, promoting, reconnaissance, and extortion identification, medicinal and logical disclosure. People have been physically extricating examples from information for quite a long time, however the expanding volume of information in current occasions has called for increasingly robotized approaches.

III. BASIC IDEA OF OUR SCHEME

As informational collections have developed in size and multifaceted nature, direct hands-on information examination has progressively been increased with circuitous, programmed information preparing. This has been helped by different revelations in software engineering, for example, neural systems, bunching, hereditary calculations, choice trees, and bolster vector machines. Information mining is the way toward applying these strategies to information with the expectation of revealing. Concealed examples. Information mining might be characterized as “the investigation and examination, via programmed or self-loader implies, of huge amounts of information so as to find significant examples and guidelines”. Thus, it might be viewed as mining learning from a lot of information since it includes learning extraction, just as information/design investigation.

Online video, a universal, visual, and exceedingly shareable medium is appropriate to intersection geographic, social, and semantic obstructions. Inclining recordings specifically, by uprightness of achieving an expansive number of watchers in a limited capacity to focus time, are incredible as both influencers and pointers of global correspondence streams. Be that as it may, are new correspondence innovations really being utilized to share thoughts all around, or would they say they are essentially reflecting previous social channels? What is more, how do social, political, and geographic variables impact global correspondence? By dissecting utilization information from computerized correspondences stages, we can start to respond to these inquiries. In this paper, we center around slanting information from the YouTube video sharing stage to look at the global utilization of Online video.

As we all know using and watching YouTube videos is a crucial a part of our everyday lives. Most people try to create their influence, income, and impact with YouTube and online video. In nutshell, most are trying to be a YouTube influencer. It will be nice if a YouTube influencer can get an idea of how the view count goes to be before making and finalizing the video. In here we tried to make a model which will help influencers to predict the amount of views for his or her next video.

IV. RELATED WORK

1) Statistical Convergence and Convergence in Statistics Authors Mark Burgin, Oktay Duman. Statistical convergence was introduced in connection with problems of series summation. The main idea of the statistical convergence of a sequence l is that the majority of elements from l converge and we do not care what is going on with other elements. We show (Section 2) that being mathematically formalized the concept of statistical convergence is directly connected to convergence of such statistical characteristics as the mean and standard deviation.

2) A Regression Approach for Prediction of Youtube Views Authors Lau Tian Rui. YouTube has grown to be the number one video streaming platform on Internet and home to millions of content creator around the globe. Predicting the potential amount of YouTube views has proven to be extremely important for helping content creator to understand what type of videos the audience prefers to watch.

3) Modelling and Statistical Analysis of Youtube's Educational Videos: A Channel Owner's Perspective Authors Samant Saurabh. YouTube is one of the most popular websites. It is a vast resource for educational content. To better understand the characteristics and impact of YouTube on education, we have analyzed a popular YouTube channel owned by the author of this paper. It has thousands of subscribers, millions of views, and hundreds of video lectures. Makers and purchasers, require opinion-mining apparatuses to gather opinions about a specific item.

4) Analysis of Youtube of Videos: A Literature Survey Authors Neha Reddy. Consumption of content from YouTube (Lanyu Shang, 2019) and other OTT (over-the-top) platforms is constantly increasing. YouTube (Lanyu Shang, 2019) being a source of education, entertainment and promotion, is a very lucrative platform. YouTubers tend to unethically attract viewers into clicking their video by manipulating their title and/or thumbnail.

5) How Youtube Developed Into A Successful Platform For User-Generated Content Authors Margaret Holland. On October 2, 2010, Felix Kjellberg uploaded a 2-minute YouTube video of himself speaking on camera while playing a video game. Today, Kjellberg, better known by his YouTube alias, “PewDiePie,”1 uploads to an online audience of over 40 million subscribers. At just 24, Kjellberg has developed his online persona into a brand name that pulls in an estimated $4 million in ad sales a year (Kain, 2014). Kjellberg is not alone.

ISSN (online): 2581-3048

Volume 8, Issue 2, pp 123-127, February-2024

https://doi.org/10.47001/IRJIET/2024.802017

V. CONCLUSION

Data mining is the way toward applying these strategies to information with the expectation of revealing concealed examples. Subsequently, it might be viewed as mining information from a lot of information since it includes learning extraction, just as information/design examination online video, a pervasive, visual, and very shareable medium, is appropriate to intersection geographic, social, and etymological boundaries. Inclining recordings specifically, by ethicalness of achieving countless in a limited ability to focus time, are groundbreaking as both influencers and pointers of global correspondence streams. Yet, are new correspondence advancements really being utilized to share thoughts all inclusive, or would they say they are just reflecting prior social channels What is more, how do social, political, and geographic variables impact global correspondence By investigating utilization information from computerized interchanges stages, we can start to respond to these inquiries. In this paper, we center around drifting information from the YouTube video sharing stage to inspect the worldwide utilization of Online video. Less relevant tutorial series or marathons can rank up their videos using targeted keywords, audience retention and video engagement. Video engagement includes sharing of video, number of subscribers, likes and views. Everything depends on number. The nature of comments is not taken into consideration.

VI. FUTURE WORK

Future work can look into In this paper select managed and unsupervised arrangement techniques. Bayesian calculations, Support Vector Machine, Deep learning, and Random Forest. In direct word classifiers (SVM, Deep Learning, Random Forest and Naïve Bayes) to develop a decision displays webbing for casting one get-together falls into the other. In other model data set passed in the different channels through find the methods. In Future our findings for measuring, analyzing, and comparing key aspects of YouTube trending videos. To know only best time to upload a video on YouTube is not enough to generate a millions of views for your Videos to become trend. There are some other factors to considered are Good Titles, Good thumbnails, Video SEO, proper tagging, and the number of subscribers are all factors that is a key generating views for your content. Understanding this Statistics will not only help YouTube to develop better algorithms to process videos but also benefit to make decisions for individual youtubers. Future project will definitely simplify the work of YouTube and content creator. By showing them their or others youtubers video analysis based on their views likes and comments from dataset what kinds of positive and negative comment has occurred on that video so user analyse them and create better content or they get motivated for uploading more videos. high efficiency as compared to the one used.

REFERENCES

[1] Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distribution. Bull. Calcutta Math. Soc. 35, 99–109 (1943).

[2] Bonacich, P.: Factoring and weighting approach to status scores and clique identification. J. Math. Soc. 2, 112–120 (1972).

[3] Borgatti, S.P., Halgin, D.S.: Analyzing Affiliation Networks, pp. 417–433. Sage (2011). Brieger, R.L.: The duality of persons and groups. Soc. Forces 53(2), 181–190 (1974).

[4] Han, J., Kamber, M.: Data Mining Concepts and Techniques (2001).

[5] Chen, N., Lin, J., Hoi, S.C., Xiao, X., Zhang, B.: Ar-miner: mining informative reviews for developers from mobile app marketplace. In: Proceedings of the 36th International Conference on Software Engineering, pp. 767–778. ACM (2014).

[6] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., Moon, S.: Analyzing the video popularity characteristics of large-scale user-generated content systems. IEEE/ACM Trans. Netw. (2009).

[7] Brodersen, A., Scellato, S., Wattenhofer, M.: Youtube around the world: geographic popularity of videos. In: Proceedings of the 21st International Conference on world Wide Web. ACM (2012).

[8] De Pauw, W., Jensen, E., Mitchell, N., Sevitsky, G., Vlissides, J., Yang, J.: Visualizing the execution of Java programs. In Software Visualization, pp. 151–162. Springer (2002).

[9] Ghorashi, S.H., Ibrahim, R., Noekhah, S., Dastjerdi, N.S.: A frequent pattern mining algorithm for feature extraction of customer reviews. Int. J. Comput. Sci. Issues (IJCSI), 29–35 (2012).

[10] Bala, S., et al.: Int. J. Comput. Sci. Mob. Comput. 3(7), 960–967 (2014).

[11] Rui, Lau & Afif, Zehan&Saedudin, Rd & Mustapha, Aida & Razali, Nazim. (2019). A regression approach for prediction of Youtube views. Bulletin of Electrical Engineering and Informatics. 8. 10.11591/eei.v8i4.1630.

[12] Pinto, Henrique & Almeida, Jussara & Gonçalves, Marcos. (2013). Using early view patterns to predict the popularity of YouTube videos. WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining 10.1145/2433396.2433443.

[13] Bärtl M. YouTube channels, uploads and views: A statistical analysis of the past 10 years. Convergence. 2018;24(1):16-32. doi:10.1177/1354856517736979.

[14] Kousha, Kayvan &Thelwall, Mike & Abdoli, Mahshid. (2012). The role of online videos in research communication: A content analysis of YouTube videos cited in academic publications. Journal of the American Society for Information Science and Technology. 63. 1710-1727. 10.1002/asi.22717.

[15] Borghol, Y, Ardon, S, Carlsson, N. (2012) The Untold Story of the Clones: Content-Agnostic Factors that Impact Youtube Video Popularity. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012, pp 1186–1194. New York: ACM.

[16] D.Nethra Pingala Suthishni, A Review on Fuzzy based packet dropping and collaborative attack detection in MANET using DSR protocol International Conference on 'Social Mobile Analytics Cloud (SMAC' 18)' organized by the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore and published in International Journal VOLUME 10.

[17] D Nethra Pingala Suthishni, KS Senthil Kumar, Fuzzy Logic Based Intrusion Detection System Using Nsl-Kdd Data Set.

[18] S. Gnanapriya, Genetic Algorithm Based Ant Colony Optimization (GA-ACO) For Cross Domain Opinion Mining, ARPN Journal of Engineering and Applied Sciences, VOL. 14, NO. 7, April 2019, ISSN 1819-6608.

Citation of this Article:

Nadhiya S, S. Dinesh Kumar, Dr. N. Sudha, Dr. G. Chitra Ganapathy, "Data Mining Method for Video Subscribers and Analysis of Youtube" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 2, pp 123-127, February 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.802017

****