

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/342150876>

Youtube Trending Video Metadata Analysis Using Machine Learning

Article · May 2020

CITATION

1

READS

2,657

5 authors, including:



[s. Amudha](#)

Dr. M.G.R. University

3 PUBLICATIONS 67 CITATIONS

[SEE PROFILE](#)



[Niveditha V.R](#)

Sathyabama Institute of Science and Technology

37 PUBLICATIONS 219 CITATIONS

[SEE PROFILE](#)



[Radha Rammohan Shanthanam](#)

Dr. M.G.R. University

21 PUBLICATIONS 88 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Health [View project](#)



Malware [View project](#)

Youtube Trending Video Metadata Analysis Using Machine Learning

S. Amudha ¹, V R. Niveditha², Dr. P.S. Raja Kumar ³, M.Revathi ⁴, Dr.S.Radha Rammohan ⁵

¹ Assistant professor Department of Computer Science and Engineering,

² Research Scholar, Department of Computer Science and Engineering,

³ Professor, Department of Computer Science and Engineering,

⁴ Assistant professor, Department of Computer Science and Engineering,

⁵ Professor, Department of Computer Applications,

^{1, 2,3,4,5} Dr.M.G.R Educational and Research Institute, Maduravoyal, Chennai-600095.

Abstract

Data Analysis and Mining are becoming indispensable part of every major organization to find recent trends and statistics and formulate business strategies, planning and marketing. However, most of the Data generated is generally in Huge Size and comes in unstructured format. Big Data cannot be analyzed by traditional database systems and processes. To resolve this issue, many new tools that implement Parallel Processing are being deployed in these organizations. As part of the Advanced Databases Project, we propose to perform Data Analysis of YouTube data. We extracted data of 5 Million Video records from YouTube API and performed Data Analysis on the data to insight into latest trends and user engagement in YouTube with respect to Categories and Year. Data Analysis and Visualization was done using Anaconda jupyter Notebook. Analysis of structured data has seen tremendous success in the past. However, analysis of large scale unstructured data in the form of video format remains a challenging area. YouTube, a Google company, has over a billion users and generates billions of views. Since YouTube data is getting created in a very huge amount and with an equally great speed, there is a huge demand to store, process and carefully study this large amount of data to make it usable.

Keywords: Data Analysis, Big Data, Mining, YouTube.

1. Introduction

Big knowledge refers to giant datasets that might not be analyzed by ancient information systems and processes like RDBMS and existing knowledge reposting systems. huge knowledge is usually characterized by Brobdingnagian Volume, High rate and High selection. firms like Google, YouTube, Facebook, Amazon, Alibaba, Pandora, and Wikipedia area unit generating and grouping Petabytes of huge knowledge each minute in mutistructured formats likes videos, audios, images, metadata, logs etc. the info generated will be used for Recommendations, formulating business and market ways exploitation knowledge Analysis and applying machine learning algorithms. it's been calculable that a pair of.5 Exabyte of information is made daily[1-10].

The Volume and form of huge knowledge makes the task of information Analysis exploitation existing ancient processing techniques very difficult. to resolve this issue, organizations area unit shifting towards exploitation multiple servers and exploitation data processing to save lots of time and memory. There area unit totally different technologies like Hadoop, Spark, HBASE that are developed and area unit speedily evolving to take care of huge knowledge.

As a part of Advanced Databases project, we've extracted and analyzed dataset of five Million records from YouTube API. The dataset size was 603 MB and consists of key attributes like Video Id, views, likes, Comments and classes. We have a tendency to performed knowledge Analysis exploitation boa Jupyter Notebook. YouTube has one.3 Billion users and three00 hours of video area unit being uploaded in YouTube each minute. YouTube gets thirty million users daily and nearly five Billion views area unit watched daily. Hence, YouTube has currently

become important promoting tool for major firms and amusement channels. the first purpose if this project is to seek out however real YouTube time knowledge will be analyzed to induce latest analysis and trends. in conjunction with videos with highest read count, most watched classes, we will area unit going to notice however the user base is increasing and the way their interests are ever-changing with each year With speedy innovations and surge of web firms like Google, Yahoo, Amazon, eBay and a speedily growing web savvy population, today's advanced systems and enterprises area unit generating knowledge in an exceedingly} very Brobdingnagian volume with nice rate and during a multi-structured formats as well as videos, images, detector knowledge, weblogs etc. from totally different sources. This has born to a replacement sort of knowledge known as huge knowledge that is unstructured someday semi structured and additionally unpredictable in nature. This knowledge is usually generated in real time from social media websites that is increasing exponentially on a each day. YouTube is one in the entire foremost fashionable and interesting social media tool and an incredible platform that reveals the community feedback through comments for printed videos, range of likes, dislikes, range of subscribers for a specific channel. YouTube collects a good form of ancient knowledge points as well as read Counts, Likes, Votes, and Comments. The analysis of the higher than listed knowledge points constitutes a really attention-grabbing knowledge supply to mine for getting implicit data regarding users, videos, classes and community interests. Most of the businesses area unit uploading their product launch on YouTube and that they uneasily look their subscribers' reviews. Major production homes launch film trailers and folks give their 1st reaction and reviews regarding the trailers. This more creates a buzz and excitement regarding the merchandise. therefore the higher than listed knowledge points become terribly crucial for the businesses in order that they will do the analysis and perceive the customers' sentiments regarding their product/services[11-20].

1.1 Scope

Our aim is to produce a scientific knowledge preprocessing analysis operating solely with the dataset US Videos. This step is important for all data processing exercises and that we wish to emphasize it. Before building theories from knowledge we'd like to grasp key knowledge attributes, like missing values, distinctive counts, outliers, and time-series trends. This kernel aims to function a tutorial to anyone fascinated by exploiting huge datasets. I focus only on the US videos dataset that isn't too huge by big-data standards (only twenty three, 362 rows by sixteen columns as of March, 2018). This knowledge set contains solely YouTube data and no data that area unit troublesome to method and store, like video, image, audio, or giant text documents. Still we are going to proceed with knowledge preprocessing and preliminary knowledge Analysis (EDA) as if this were a very huge dataset, using techniques that might be utilized in rather more difficult knowledge manning exercises. We have a tendency to worker variety of techniques from the Scikit/Learn toolkit to administer aspiring to the info at hand.

1.2 Proposed Best Practices

1. Spot knowledge errors or poor knowledge quality as presently as doable. I confer with errors that tend to arise out of issues in knowledge assortment, knowledge handling, or knowledge transfer. This can be not being a giant drawback here.

2. For knowledge given in time snapshots spot periods of no variation or otherwise sudden variation. Investigate if key parameters drift in time. Our dataset, us videos, will show a lot of higher values for views, likes, associated variables in 2018 than in 2017. Hence, it's risky to extrapolate calculable relationships supported 2017 values.

3. Avoid unnecessarily difficult process throughout knowledge visual image. This tends to occur with numerical variables of very huge datasets. Scatterplots from a billion row dataset will take a really very long time to end and don't seem to be simple to scan. we are going to utilize binning for these tasks (histograms for example bin 1st and so plot) or we have a tendency to plot knowledge visuals employing a random sample. good plotting modules tend to try to this within the background.

4. Spot variables that require preprocessing, e.g. reformatting, filtering out. We discover that rendering exploitation trending date and video_id helps the reader to grasp the structure of the dataset, us videos.

1.3 Data mining Techniques

1. Tracking patterns. One in all the foremost basic techniques in data processing is learning to acknowledge patterns in your knowledge sets. This can be typically a recognition of some aberration in your knowledge happening at regular intervals, or Associate in Nursing ebb and flow of a definite variable over time. As an example, you would possibly see that your sales of a definite product appear to spike simply before the vacations, or notice that hotter weather drives additional individuals to your web site.

2. Classification. Classification may be a additional complicated data processing technique that forces you to gather varied attributes along into discernible classes, that you'll be able to then use to draw more conclusions, or serve some operate. as an example, if you're evaluating knowledge on individual customers' money backgrounds and get histories, you would possibly be able to classify them as "low," "medium," or "high" credit risks. You'll then use these classifications to find out even additional regarding those customers.

3. Association. Association is expounded to following patterns, however is additional specific to dependently connected variables. during this case, you'll explore for specific events or attributes that area unit extremely related with another event or attribute; as an example, you would possibly notice that once your customers get a selected item, they additionally typically get a second, connected item. this can be typically what's wont to populate "people additionally bought" sections of on-line stores.

1.4. Data Preparation

The knowledge preparation method consumes regarding ninetieth of the time of the project. The knowledge from totally different sources ought to be designated, cleaned, remodeled, formatted, anonym zed, and created (if required). Data cleanup may be a method to "clean" the knowledge the info the information by smoothing buzzing data and filling in missing values. For example, for a client demographics profile, age knowledge is missing. The info is incomplete and may be crammed. In some cases, there may well be knowledge outliers. For example, age includes a price three hundred. knowledge may well be inconsistent. For example, name of the client is totally different in several tables. Data transformation operations amendment the info to create it helpful in data processing. Following transformation will be applied.

1.5 Data Cleanup

```
print(us_videos[us_videos.video_error_or_removed])
us_videos = us_videos[~us_videos.video_error_or_removed]

title \

trending_date video_id
2017-11-25 RK_B4Ez4_5Q Verizon 360 Live: The Macy's Thanksgiving Day ...
2018-02-01 kZete48ZtsY Deleted video
2018-02-02 kZete48ZtsY Deleted video
2018-02-03 kZete48ZtsY Deleted video
channel_title category_id \
trending_date video_id
2017-11-25 RK_B4Ez4_5Q Verizon 24
2018-02-01 kZete48ZtsY Midnight Video 1
```

2018-02-02	kZete48ZtsY	DaHoopSpot Productions	17
2018-02-03	kZete48ZtsY	DaHoopSpot Productions	17

tags \

trending_date video_id

2017-11-25	RK_B4Ez4_5Q	live stream "360 video" "fun videos for kids" ...
2018-02-01	kZete48ZtsY	horror "horror short" "short" "short film" "my...
2018-02-02	kZete48ZtsY	[none]
2018-02-03	kZete48ZtsY	[none]

views likes dislikes comment_count \

trending_date video_id

2017-11-25	RK_B4Ez4_5Q	2618344	45197	2315	3332
2018-02-01	kZete48ZtsY	60262	4804	122	736
2018-02-02	kZete48ZtsY	2611	8	12	5
2018-02-03	kZete48ZtsY	2620	8	12	5

thumbnail_link \

trending_date video_id

2017-11-25	RK_B4Ez4_5Q	https://i.ytimg.com/vi/RK_B4Ez4_5Q/default.jpg
2018-02-01	kZete48ZtsY	https://i.ytimg.com/vi/-V1Oo7srGf0/default.jpg
2018-02-02	kZete48ZtsY	https://i.ytimg.com/vi/NzCSJrxQyQI/default.jpg
2018-02-03	kZete48ZtsY	https://i.ytimg.com/vi/NzCSJrxQyQI/default.jpg

comments_disabled ratings_disabled \

trending_date video_id

2017-11-25	RK_B4Ez4_5Q	False	False
2018-02-01	kZete48ZtsY	False	False
2018-02-02	kZete48ZtsY	False	False
2018-02-03	kZete48ZtsY	False	False

video_error_or_removed \

trending_date video_id

2017-11-25	RK_B4Ez4_5Q	True
2018-02-01	kZete48ZtsY	True
2018-02-02	kZete48ZtsY	True
2018-02-03	kZete48ZtsY	True

description \

trending_date video_id

2017-11-25	RK_B4Ez4_5Q	This year, we hid special offers throughout th...
2018-02-01	kZete48ZtsY	After dusting off an old Mystery Date-style bo...
2018-02-02	kZete48ZtsY	NaN

```
2018-02-03 kZete48ZtsY NaN
publish_date days_to_trending dislike_percentage
trending_date video_id
2017-11-25 RK_B4Ez4_5Q 2017-11-23 2 0.048725
2018-02-01 kZete48ZtsY 2018-01-29 3 0.024767
2018-02-02 kZete48ZtsY 2017-12-16 48 0.600000
2018-02-03 kZete48ZtsY 2017-12-16 49 0.600000
```

- The Thanksgiving Day Parade is not an obvious candidate for deletion.
- It's interesting that live events are included in the dataset. These are likely to trend immediately and lose their popularity within a couple of days.

2. Project Description

This paper apply machine learning classification algorithm to determine the efficient courier from the given data set. In this project the prediction of the courier services is done in R-Studios with the help of decision tree algorithm.

2.1 Video Category Distribution

```
sns.set(font_scale=1.5,rc={'figure.figsize':(11.7,8.27)})
sns_ax = sns.countplot([categories[i] for i in us_videos.category_id])
_, labels = plt.xticks()
_ = sns_ax.set_xticklabels(labels, rotation=60)
```






Photo	Channel Name	Title	Category	Publish Date	Days Trending	Views
	Jack's films	"cough"	Comedy	2018-02-26	16	1,126,368
	Lucas the Spider	Lucas the Spider - Polar Bear	Film & Animation	2019-03-04	16	613,805
	ChrisYoungVEVO	Chris Young - Hangin' On	Music	2018-02-26	16	76,832
	Walt Disney Animation Studios	Ralph Breaks The Internet: Wreck-It Ralph 2 Official Teaser Trailer	Film & Animation	2018-02-28	15	4,223,613
	MeghanTrainorVEVO	Meghan Trainor - No Excuses	Music	2019-03-01	15	1,034,355

Fig 1 Video Category Distribution

2.2 Time Variation Test

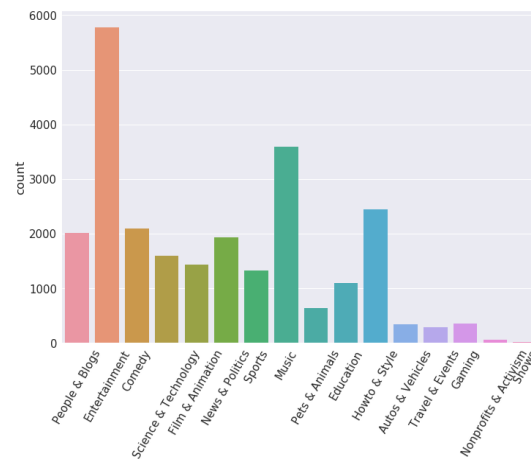


Fig 2 Time Variation Test

```
table = pd.pivot_table(us_videos, index=us_videos.index.labels[0])
table.index = us_videos.index.levels[0]
_ = table[['likes','dislikes','comment_count']].plot()
_ = table[['views']].plot()
_ = table[['comments_disabled','ratings_disabled','video_error_or_removed']].plot()
```

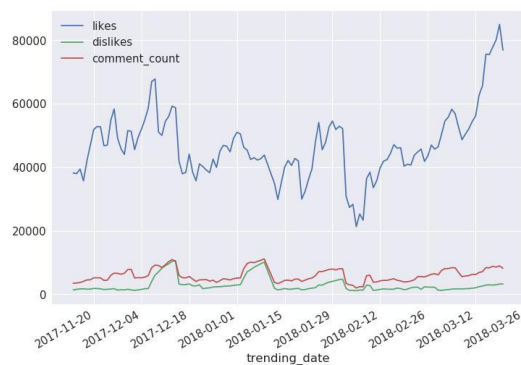


Fig 3 Trending Data Like, Dislike, Comment

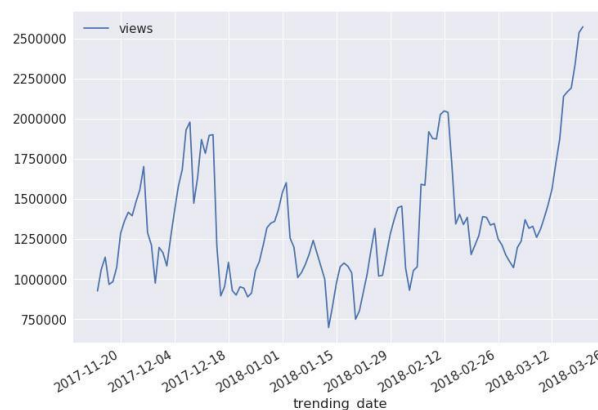


Fig 4 Trending by Views

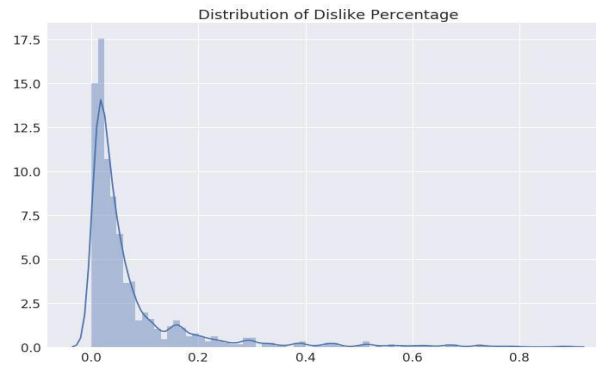


Fig 5 Distribution of Dislike Percentage

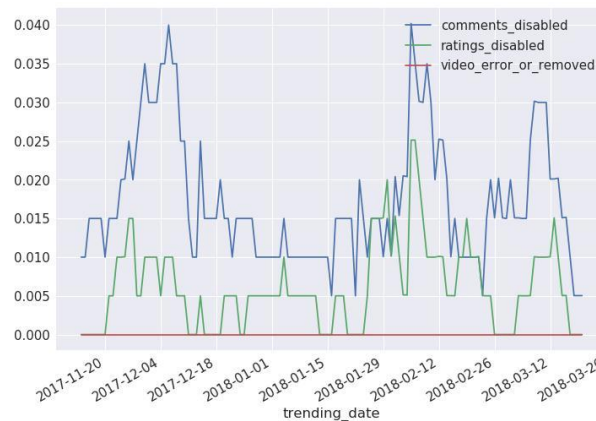


Fig 6 Trending by Comments, Ratings, Video Error or Removed

2.3 Data mining

```

tmp = video_level[(video_level.views_ratio < 10) & (video_level.freq > 1) &
(video_level.views_ratio > .8)].dropna().sort_values(by='views_ratio')

cat_ratio_median = tmp.groupby('category')['views_ratio'].median()

tmp = tmp.merge(cat_ratio_median.rename('cat_ratio_median').to_frame(),
left_on='category',right_index=True)

y = np.log(tmp.views_ratio)

print('y')

print(y.describe(percentiles=[.05,.25,.5,.75,.95]))

X =
tmp[['views','likes','dislikes','comment_count','dislike_percentage','days_to_trending','cat_r
atio_median']]

tmp_log =
np.log(tmp[['views','likes','dislikes','comment_count','dislike_percentage','days_to_trending
','cat_ratio_median','views_ratio']]+1)

X_reg =
tmp_log[['views','likes','dislikes','comment_count','dislike_percentage','days_to_trending','c
at_ratio_median']]

y_reg = tmp_log.views_ratio

print('y_reg')

print(y_reg.describe(percentiles=[.05,.25,.5,.75,.95]))

```

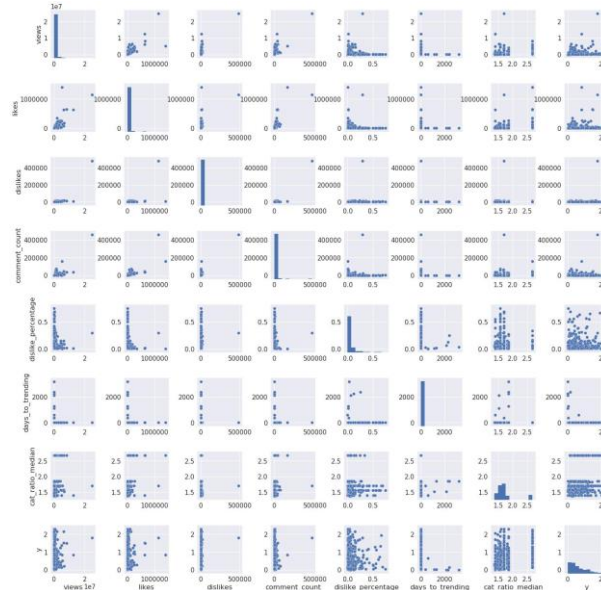
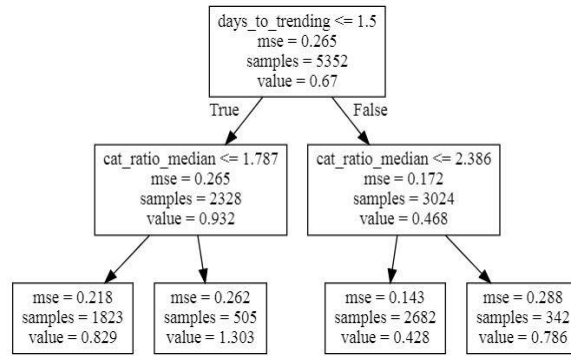



Fig 7 Views,Likes,Dislikes,Comments,etc-1

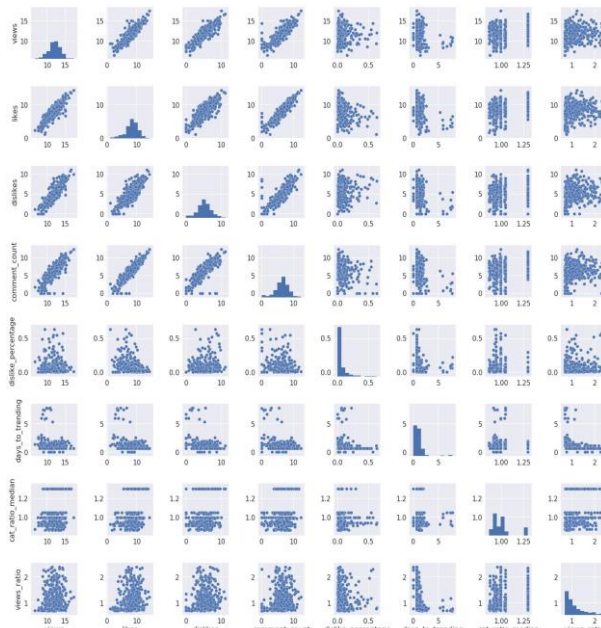


Fig 7 Views, Likes, Dislikes, Comments,etc-2

3. Conclusion

The future work would come with extending the analysis of YouTube knowledge exploitation different huge knowledge analysis Technologies like Pig and Map Reduce and do a feature comparison analysis. It'd be attention-grabbing to ascertain that technology fares higher as compared to the opposite ones.

One feature that's not extra within the project is to represent the output during a Graphical computer programmer (GUI). the present project displays a really oversimplified output that doesn't warrant a user interface. However, if the output is simply too giant and sophisticated, the output will be interfaced during a user interface format to show the results. the info will then be given in several format as well as pie-charts and graphs for higher user expertise.

Another doable extension of this project may well be the YouTube Comment Analysis project. The present scope of the project includes analyzing the statistics for a channel/category as well as read counts, likes, dislikes, country wise read etc. By distinguishing classifying/categorizing the polarity of the words, sentiment analysis or opinion minding will be performed for a selected video. This might tell US writer's perspective towards a specific product or a given subject. Exploitation Sentiment Analysis, we are able to verify if the overall perspective of individuals is positive, negative or neutral towards a selected subject/video

References

- [1] H. Li, X. Cheng, and J. Liu, "Understanding video sharing propagation in social networks: Measurement and analysis," *ACM Trans. Multimed. Comput. Commun. Appl. TOMM*, vol. 10, no. 4, p. 33, 2014.
- [2] V. R. Niveditha et.al, Detect and Classify Zero Day Malware Efficiently In Big Data Platform, *International Journal of Advanced Science and Technology*, 29(4s), 2020, 1947-1954.
- [3] V. R. Niveditha and Ananthan TV, "Improving Acknowledgement in Android Application", *Journal of Computational and Theoretical Nano science*. 16, (2019), pp. 2104–2107.
- [4] Natrayan, L., and M. Senthil Kumar. "A potential review on influence of process parameter and effect of reinforcement on mechanical and tribological behaviour of HMMC using squeeze casting method". *Journal of Critical Reviews*, Vol 7, Issue 2, (2020), pp.1-5.
- [5] Natrayan, L and M. Senthil Kumar. Influence of silicon carbide on tribological behaviour of AA2024/Al₂O₃/SiC/Gr hybrid metal matrix squeeze cast composite using Taguchi technique." *Mater. Res. Express*, 6, (2020), pp.1265f9.
- [6] Dahlia Sam et al., "Progressed IOT Based Remote Health Monitoring System", *International Journal of Control and Automation*, 13(2s), (2020), pp. 268-273.
- [7] L. Natrayan, M. Senthil Kumar, and M. Chaudhari, Optimization of Squeeze Casting Process Parameters to Investigate the Mechanical Properties of AA6061/Al₂O₃/SiC Hybrid Metal Matrix Composites by Taguchi and Anova Approach. *Advances in Intelligent Systems and Computing*, 949, (2020), pp.393-4062020
- [8] P.Sakthi Shunmuga Sundaram et al. "Smart Clothes with Bio-sensors for ECG Monitoring", *International Journal of Innovative Technology and Exploring Engineering*, Volume 8, Issue 4, (2019), pp. 298-30.
- [9] S. Velliangiri, P. Karthikeyan & V. Vinoth Kumar (2020) Detection of distributed denial of service attack in cloud computing using the optimization-based deep networks, *Journal of Experimental & Theoretical Artificial Intelligence*, DOI: 10.1080/0952813X.2020.1744196

- [10] Praveen Sundar, P.V., Ranjith, D., Vinoth Kumar, V. et al. Low power area efficient adaptive FIR filter for hearing aids using distributed arithmetic architecture. *Int J Speech Technol* (2020). <https://doi.org/10.1007/s10772-020-09686-y>
- [11] Vinoth Kumar V, Karthikeyan T, Praveen Sundar P V, Magesh G, Balajee J.M. (2020). A Quantum Approach in LiFi Security using Quantum Key Distribution. *International Journal of Advanced Science and Technology*, 29(6s), 2345-2354.
- [12] Umamaheswaran, S., Lakshmanan, R., Vinothkumar, V. et al. New and robust composite micro structure descriptor (CMSD) for CBIR. *International Journal of Speech Technology* (2019), doi:10.1007/s10772-019-09663-0
- [13] Karthikeyan, T., Sekaran, K., Ranjith, D., Vinoth kumar, V., Balajee, J.M. (2019) "Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques", *International Journal of Web Portals (IJWP)*, 11(2), pp.41-52
- [14] Vinoth Kumar, V., Arvind, K.S., Umamaheswaran, S., Suganya, K.S (2019), "Hierarchal Trust Certificate Distribution using Distributed CA in MANET", *International Journal of Innovative Technology and Exploring Engineering*, 8(10), pp. 2521-2524
- [15] Maithili, K , Vinothkumar, V, Latha, P (2018). "Analyzing the security mechanisms to prevent unauthorized access in cloud and network security" *Journal of Computational and Theoretical Nanoscience*, Vol.15, pp.2059-2063.
- [16] V.Vinoth Kumar, Ramamoorthy S (2017), "A Novel method of gateway selection to improve throughput performance in MANET", *Journal of Advanced Research in Dynamical and Control Systems*,9(Special Issue 16), pp. 420-432
- [17] Dhilip Kumar V, Vinoth Kumar V, Kandar D (2018), "Data Transmission Between Dedicated Short-Range Communication and WiMAX for Efficient Vehicular Communication" *Journal of Computational and Theoretical Nanoscience*, Vol.15, No.8, pp.2649-2654
- [18] Kouser, R.R., Manikandan, T., Kumar, V.V (2018), "Heart disease prediction system using artificial neural network, radial basis function and case based reasoning" *Journal of Computational and Theoretical Nanoscience*, 15, pp. 2810-2817
- [19] Shalini A, Jayasuruthi L, Vinoth Kumar V, "Voice Recognition Robot Control using Android Device" *Journal of Computational and Theoretical Nanoscience*, 15(6-7), pp. 2197-2201
- [20] Jayasuruthi L, Shalini A, Vinoth Kumar V., (2018) " Application of rough set theory in data mining market analysis using rough sets data explorer" *Journal of Computational and Theoretical Nanoscience*, 15(6-7), pp. 2126-2130