# “Feature trend forecasting using text data”

*by* Kavi Shankar K S

**Submission date:** 20-May-2021 05:18PM (UTC+0530)
**Submission ID:** 1590261604
**File name:** Capstone_phase_1_review_3_report.pdf (2.39M)
**Word count:** 4860
**Character count:** 25936


**UE18CS390A**

# “Feature trend forecasting using text data”

Kavishankar K S PES1201802001

Mohammed Zeeshan PES1201801814

Rohan M PES1201801798

Under the guidance of

## Prof.Dinesh Singh

Designation
PES University

January - May 2021

# ABSTRACT

We are in the age of the Internet , where it has grown like never before. It was impossible to even imagine how easy the internet has made all our living styles and also how powerful and impactful it has become in all our lives. Social media and platforms like Google , YouTube , Twitter have undertaken the part of becoming a powerful tool to connect people even on a global level like never before . People always want to stay updated with all present live news.To keep track of all that we need to forecast the most possible trend that could happen.

Internet and Social media are very powerful platforms nowadays , there are about 3.9 billion active social media users and 4.6 active Internet users. People get easily influenced by using them in aspects like socially , economically , politically and can also result in huge events around us. These can only be predicted by actively following the latest trends towards which people are interested. So, this shows how important it is to Forecast the trends which might occur in the near future.

There are many trends like trend components, seasonal component, cyclical component and irregular component but among these trend components irregular components cannot be forecasted because they don't have any significant data. Here we go with approach mining of data from social media, and we proposed a method of gathering data from social media websites and then extensively pre-processing data using text preprocessing techniques such as NLP technique of normalisation, stop word removal, stemming and lemmatization. Then correlation between each topic is analysed after the classification of text into the respective domains has been done. Then the classifieds are subjected to forecasting using the neural network model and then presenting the results to the client in the web interface.

Finding the trending topics from various social media platform like twitter and youtube which provides us the daily trending topics based on the number of times a particular topics are pronounced in their platform and google trends api provides us trends with respect to topics and number of searched made for that topic based on the keyword query to google from multiple geographic location by various content browsers and internet users.

# TABLE OF CONTENT

# 1) CHAPTER 1

## INTR0DUCTI0N

---

Growth in the Internet has brought a very large and even growing popularity among content creators . platforms like YouTube is often considered as one among the top 3 most popular and liked web applications for users and content creators. The quantity of data which is on YouTube in 60 days is equal to the amount of data transmitted for 720 months by ABC , NBC and CBS together . This shows the content upload rate, This shows that how the contents are undistributed in YouTube. Reports also show that almost 72hours of videos are content every minute . This shows how important it is to forecast future trends.

Internet and Social media are very powerful platforms nowadays , there are about 3.9 billion active social media users and 4.6 active Internet users. People get easily influenced by using them in aspects like socially , economically , politically and can also result in huge events around us. These can only be predicted by actively following the latest trends towards which people are interested. So, this shows how important it is to Forecast the trends which might occur in the near future.

Forecasting vastly used in the prescriptive analysis for predicting the contents of what will happen in the future. Demand forecasting has become a massive forecasting application where people analyse the contents of present and past to forecast the future. Fashion forecasting is also done to predict the future fashion trends. Researchers also actively forecasted Cholera outbreaks by successfully analysing large data of newspapers and articles.

In this condition , Forecasting content popularity will be of greater significance to support and drive various design and management of services. Let's assume that , e-marketing, in planning advertising campaigns and estimating costs can be done by using information which is forecasted the future popular type of content. For an effective information service accurately predicting content popularity is the key property .

Content forecasting can improve quality of services and help us to know what content is going to be the most possible trending topics for users and content creators using various techniques.

# 2) CHAPTER 2

## Problem statement

---

Discovering trending topics from social media platforms and Forecasting the future trends based on the recent and past trend of the topics collected.

Content creators find it very difficult in finding the perfect topics on which they can create a trending content. Content creating platforms nowadays is also a very high revenue generating resource and many consider it also a career and professional way. This requires proper guidance for beginning content creators and they might find it difficult to cope up with a very competitive social media environment.

Trend is movement in which component changes may be upward and downward can be developing or

changing.

There are 4 types of trends mainly:

1. Trend Component - consistent upward and downward movement of data over time period
2. Cyclical trend -occurs within cycle period of time
3. Season trend- Occurs within a calendar year of time
4. Irregular trend- white noise uncorrelated trend.

So finally forecasting the trend is a major issue that has to be dealt with using the good measures to address the problems of forecasting.

Also there are other aspects of these projects like in Social analysis where nowadays people are influenced by social media like Youtube , Twitter and Google in an enormous way which can lead to a very significant physiological aspect of people using them. So this can only be controlled or prevented if we know the future changing trends that have a chance to occur around us.

Trends not only affect the social aspect of people , it also affects financial and marketing functions.Users will quickly get influenced by social media and this affects large industrial sectors , stock market prices also. So, if you know the flow of trends then only we can make good decisions in the market , stocks and in society.

## 2.1 Overview

Forecasting the future trend it may be a downward or upward trend of the topic using the extracted text data from social media platforms.

## 2.2 Objective

Identifying the procedure for extracting the required data and preprocessing the extracted data and classification of the data which we are gathered to forecast using the machine learning model to find out the trend of the topics and interpreting the outcome of the project.

## 2.3 Scope

In the many field researchers are really very curious about finding out the forthcoming topics i.e. finding out the future. But, actually, the future is nothing but a consequence of the present and past so by analysing the present and past data in a very effective and careful way is a really tedious task by human interpretation. So, we are using the method of machine learning to make this task easier ad more accurate than the humans doing it manually.

Helps for the business decision R&D part of the company to predict the future trends.For content creators on youtube or writing blogs or posts. Decisions which are made based on upcoming trends. Products that are customised based on trend. E-commerce websites can build their stock according to trend.

Demand forecasting is often the main advantage of forecasting that helps to tackle and manage many industries to compensate for the upcoming demands of its product where they maintain their SKU (stop keeping unit) ready for the future to address the problems.

# 3) CHAPTER 3

## LITERATURE SURVEY

---

## 3.1 : "Topic discovery and future trend forecasting for texts."

**Page Source:**Click Here

In this, researchers presented the method for extraction of trending topics from the corpus of research documents and technical data and applications patents and forecast the data using models.

**Data used**: Corpus of research and patents publications.

**Proposed Methodology:**

The proposed methodology contains the methods for the extraction and pre-process data which are converted into individual topic transactions for forecasting and analysis of correlation between each topic. Finally forecasting data using a regression model provided by some plugin.

**Drawbacks:**

They used a very simple forecasting model which is less efficient in prediction.

Pre-processing can be extended to some extent by using stemming and stopping.

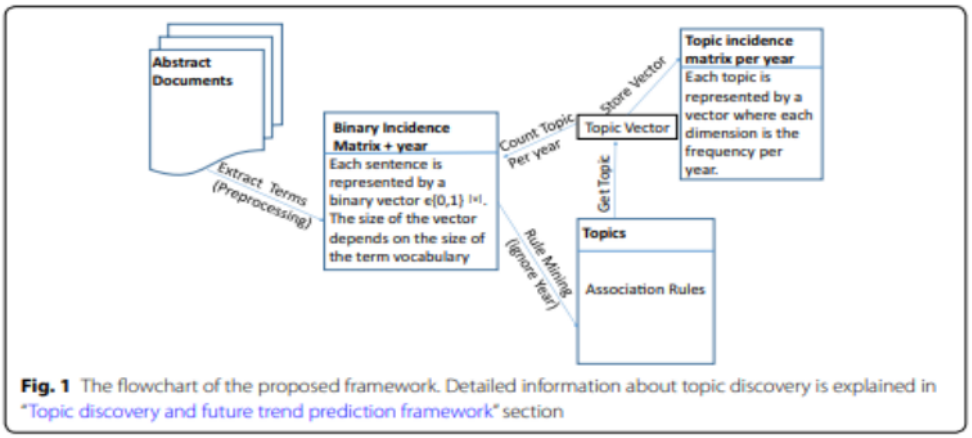



**Fig. 1** The flowchart of the proposed framework. Detailed information about topic discovery is explained in "Topic discovery and future trend prediction framework" section

3.1.1 Paper discussion:

They have used association analysis processes on data to extract the topics list and then the temporal correlation analysis for analysing the correlation between individual topics to measure the extent of correlation between the topics and it also helps in discovering networks. Finally, forecasting is used for prediction of the popular topic.

3.1.2 Steps

**1) Data extraction**:

Inclusively data is extracted from the corpus of research papers and various platforms. They have used sentence-level association rule mining to collect the transaction in the collected dataset.

**2) Pre-process:**

Each extracted sentence is converted into transactions and the keywords in transactions are considered as items in transactions.

**3)Forecasting**

Finally, the pre-processed dataset is forecasted using a plugin called WEKA which helps to use many models for prediction but in this they have used regression for prediction of the probability of the topic may trend in the future.

## 3.2 Paper 2:

### "FORECASTING TIME SERIES DATA USING HYBRID GREY RELATIONAL ARTIFICIAL NEURAL NETWORK AND AUTO REGRESSIVE INTEGRATED MOVING AVERAGE MODEL"

A new hybrid model of gran_arima is proposed for dealing with time variate linear and nok linear model.

#### 3.2.1 Paper discussion.

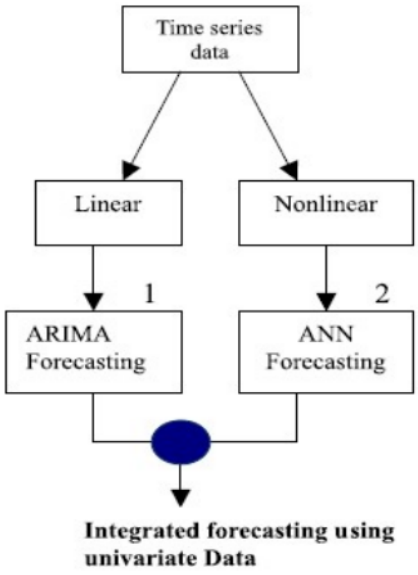
Here authors have proposed to use a hybrid model by combining grey relational artificial neural network and auto regressive integrated moving average ,

Hybrid models are good in time variate data sets and handles linear and non linear data efficient than individual models. the proposed hybrid model is more efficient than non hybrid models and can be used with a small data set also, we are thinking of using this hybrid model in our project we can make some changes to tweak this model for our project ,The changes to be made will be decided in a later phase of capstone.

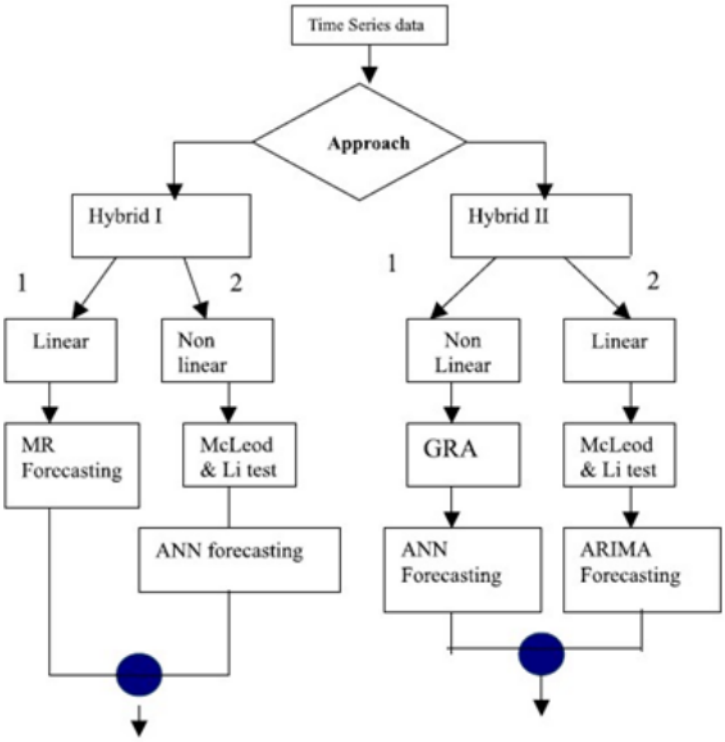
Forecasting value of grann_arima comparatively accurate as it can handle both linear and non-linear patterns in time series data.

· grann_arima also performs well in small datasets..

## Conventional model:



**New Hybrid model:**




### 3.2.2 Steps

### 1) Data extraction:

KLSE time variate DATA is used for demonstration in paper.

This contains data relating to stock exchange which is a time variate data.

### 2) Pre-process

KLSE Data is subdivided into two following parts:

1. test data

2.training data.

9

The test data is used for getting the accuracy of the model on non trained data.

**3)Forecasting**

Finally, the preprocessed dataset is forecasted for prediction of the probability of the topic may trend in the future.

## 3.3 Paper 3 :

## " Using Early View Patterns to Predict the Popularity of YouTube Videos."

Forecasting content popularity will be helpful in various services like recommendation systems, advertising, finances , markets , social analysis etc..

This paper shows the **baseline model** leads to a decrease in relative square errors by 20%.

### Prediction models and methods to be used :

Prediction models are used to get the most feasible future contents.

Terms used are :

1. total number of views of a video at **tt**
2. data from the rst **tr** days

where (tr < tt).

1. **Relative Squared Error :**

   Helps to evaluate performance of ML models

   N(v; t) -> total number of views video v receives up to day

   N (v; tr; tt) -> total number of views predicted for v in date tt data from the rst tr days

   RSE prediction is :

$$RSE = \left( \frac{\hat{N}(v, t_r, t_t)}{N(v, t_t)} - 1 \right)^2$$

C -> collection of videos

We get

$$mRSE = \frac{1}{|C|} \cdot \sum_{v \in C} \left( \frac{\hat{N}(v, t_r, t_t)}{N(v, t_t)} - 1 \right)^2$$

## 2. Szabo-Huberman Model (S-H Model):

Szabo and Huberman expresses the future popularity of a content v as:

$$\hat{N}(v, t_r, t_t) = \alpha_{t_r, t_t} \cdot N(v, t_r)$$

we can compute the optimal value to a given pair (tr; tt)

Let c -> set of training data videos

by plugging the expression for ^N(v; tr; tt) in taking its derivative and equating it to zero. This leads to:

$$\alpha_{t_r, t_t} = \frac{\sum_{v \in C} \frac{N(v, t_r)}{N(v, t_t)}}{\sum_{v \in C} \left( \frac{N(v, t_r)}{N(v, t_t)} \right)^2}$$

Optimal value for parameter model can be computed within 0(n) for n training data.

3. **Multivariate Linear Model (ML Model) :**

**This model predicts the trending of a video at tt as a linear function . This linear assumption is very reasonable for the strong linear correlation between past and future popularities observed in .**

More formally, let $x_i(v)$ be the number of views received by video $v$ on the $i$-th day since its upload ($x_i(v) = N(v,i) - N(v,i-1)$). The *feature vector* $X_{t_r}(v)$ is defined as:

$$X_{t_r}(v) = (x_1(v), x_2(v), \ldots, x_{t_r}(v))^T$$

and we estimate the popularity of the video $v$ at $t_t$ as:

$$\hat{N}(v, t_r, t_t) = \Theta_{(t_r,t_t)} \cdot X_{t_r}(v) \quad (5)$$

where $\Theta_{(t_r,t_t)} = (\theta_1, \theta_2, \ldots, \theta_{t_r})$ is the vector of *model parameters* and depends only on $t_r$ and $t_t$.

Given a training set $C$, $t_r$ and $t_t$, we can compute the optimal values for the elements of $\Theta_{(t_r,t_t)}$ as the ones that minimize the mRSE on $C$, i.e.:

$$\arg\min_{\Theta_{(t_r,t_t)}} \frac{1}{|C|} \sum_{v \in C} \left( \frac{\Theta_{(t_r,t_t)} \cdot X_{t_r}(v)}{N(v,t_t)} - 1 \right)^2 \quad (6)$$

The hypothesis is that $\hat{N}(v, t_r, t_t)$ is a linear function, and $N(v, t_t)$ a scalar. Thus, it follows that:

$$\frac{\Theta_{(t_r,t_t)} \cdot X_{t_r}(v)}{N(v,t_t)} = \Theta_{(t_r,t_t)} \cdot \left( \frac{X_{t_r}(v)}{N(v,t_t)} \right)$$

Let $X_v^* = \frac{X_{t_r}(v)}{N(v,t_t)}$. We express the optimization problem as:

$$\arg\min_{\Theta_{(t_r,t_t)}} \frac{1}{|C|} \sum_{v \in C} \left( \Theta_{(t_r,t_t)} \cdot X_v^* - 1 \right)^2, \quad (7)$$

**ordinary least squares (OLS) problem using the singular value decomposition can predict within with complexity O(np2),**

**where n -> number of training data.**

4. **MRBF Model :**

Radial Basis Functions (RBFs) are used with MRBF Model to find similarities between data which helps us to reduce prediction errors in a very significant way.

$$\hat{N}(v, t_r, t_t) = \underbrace{\Theta_{(t_r,t_t)} \cdot X(v)}_{\text{ML Model}} + \underbrace{\sum_{v_c \in C} \omega_{v_c} \cdot RBF_{v_c}(v)}_{\text{RBF Features}}$$

Where,

C -> set of training set chosen

Wvc -> model weight associated with the RBF

# 4) CHAPTER 4

# DATA

---

## 4.1 Overview

In order to forecast trends we get data from 3 different sources . This helps us to reduce the prediction errors .

Sources are :

1. Youtube
2. Tweeter
3. Google trends

## 4.2 Data from YouTube

Data collected from YouTube are of two categories .

They are :
1. Past youtube trends
2. Live youtube trends

### 4.2.1 Past youtube trends

Dataset has data from past several months old daily live YouTube trends data.Data set included are from different countries which are represented by their short forms.

The headers in the video file are:

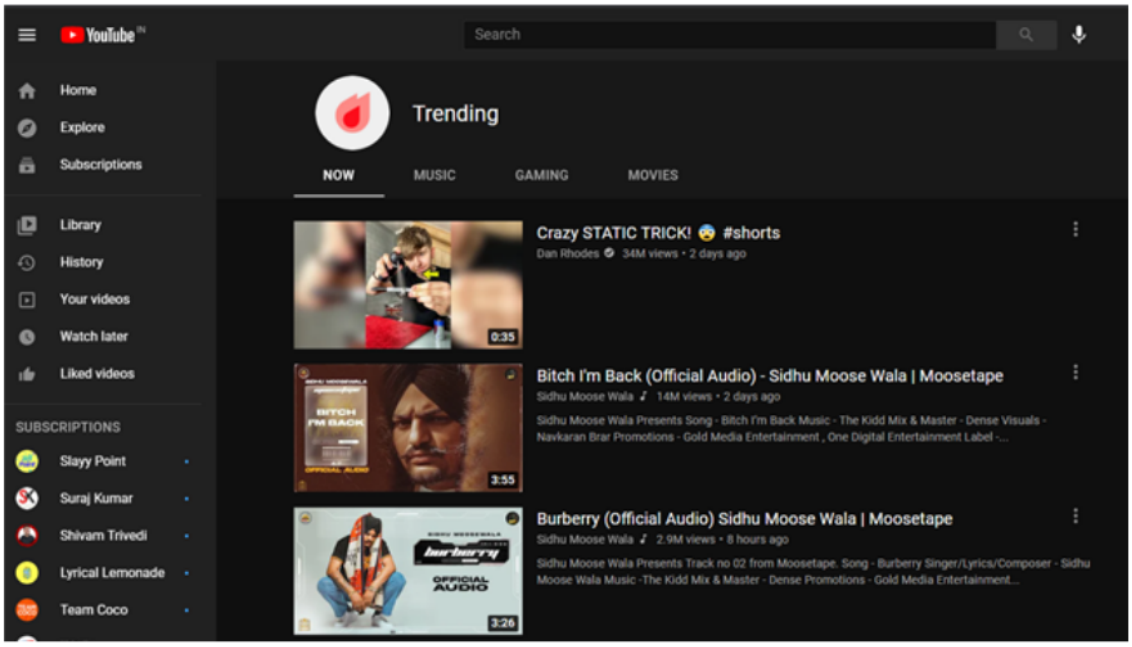
- video_id (Common id field to both comment and video csv files)
- title
- channel_title
- category_id (Can be looked up using the included JSON files, but varies per region so use the appropriate JSON file for the CSV file's country)
- tags (Separated by | character, [none] is displayed if there are no tags)
- views
- likes
- dislikes
- thumbnail_link
- date (Formatted like so: [day].[month])

### 4.2.2 Live youtube trends

We use YouTube API and python libraries service which gives us the live YouTube trends in the form of csv file,

Sample image output:

| | video_id | title | publishedA | channelId | channelTit | categoryId | trending_d | tags | view_cour | likes | dislikes | c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | i0Ye1lBEgr | Crazy STATIC TRICK! ðŸ˜¨ #shorts | 2021-05-1 | UC6D1L2v | Dan Rhode | 24 | 21.17.05 | [none] | 34610277 | 1421177 | 49412 | |
| 3 | W5NgXKe | Bitch I'm Back (Official Audio) - Sic | 2021-05-1 | UC9ChdqC | Sidhu Moc | 10 | 21.17.05 | sidhu moo | 14054753 | 953832 | 34888 | |
| 4 | h8r0z_q--t | Burberry (Official Audio) Sidhu Mo | 2021-05-1 | UC9ChdqC | Sidhu Moc | 10 | 21.17.05 | sidhu moo | 2924581 | 506554 | 8639 | |
| 5 | utvmZJDyC | Tauktae Cyclone à¤•à¤¾à¤µà¤¾ à¤ | 2021-05-1 | UCYUPBf8 | BBC News | 22 | 21.17.05 | bbc gujara | 531736 | 4186 | 294 | |
| 6 | HiBwVq10 | 30 Weds 21 Web Series \| Episode | 2021-05-1 | UCfrNV2iJ | Girl Formu | 24 | 21.17.05 | Girl Formu | 3059272 | 128285 | 4169 | |
| 7 | cYOkgHzsI | #shorts | 2021-05-1 | UCGLDtG2 | Tool_Tips | 22 | 21.17.05 | [none] | 922206 | 18331 | 2599 | |
| 8 | 29xpBEEJS | BATTLEGROUNDS MOBILE INDIA - | 2021-05-1 | UCe31NPE | BATTLEGR | 20 | 21.17.05 | [none] | 9230313 | 920960 | 12560 | |
| 9 | xkUyhYzSS | Israel Palestine Conflict: à¤‡à¤¸à¤ | 2021-05-1 | UCN7B-QI | BBC News | 25 | 21.17.05 | BBC Hindi\| | 815779 | 0 | 0 | |
| 10 | 0E-vPTeiol | Ajjubhai94 and Amitbhai VS Rome | 2021-05-1 | UC5c9VlY1 | Total Gam | 20 | 21.17.05 | free fire\|g | 3333620 | 447768 | 9609 | |
| 11 | YQNL4Og2 | Papa Ki Pari #shorts | 2021-05-1 | UCBZQAxI | Yashraj Mu | 10 | 21.17.05 | Papa ki pa | 3369776 | 462792 | 7656 | |




## 4.3 Data from Twitter

We use Twitter Streaming API to continuously extract tweets from 00:05 to 23:55 local time. The Streaming API provides a stream of tweets according to filters provided by the user. Twitter only provides 1% of the total tweets filtered and Twitter also provides the Twitter IDs of the missed tweets as part of the stream.

We also use the **Latent Dirichlet Allocation algorithm** to create a word cloud which represents the trending topics.

```
Hashtag: Daniel Kaluuya       volume of tweets: 120587
Hashtag: Harrison Ford       volume of tweets: 10539
Hashtag: Trent Reznor       volume of tweets: 10283
Hashtag: Jessica       volume of tweets: 63528
Hashtag: Sound of Metal       volume of tweets: 48492
Hashtag: Don Sergio       volume of tweets: 48411
Hashtag: Chloé Zhao       volume of tweets: 125131
Hashtag: Globo       volume of tweets: 116207
Hashtag: □ □ □ □       volume of tweets: None
Hashtag: bbb18       volume of tweets: 33031
Hashtag: Corinthians       volume of tweets: 88744
Hashtag: Pocah e Arthur       volume of tweets: 63646
Hashtag: Fernando Carlos       volume of tweets: None
Hashtag: Kershaw       volume of tweets: None
Hashtag: Steven Yeun       volume of tweets: 20134
Hashtag: Aranza       volume of tweets: 16211
Hashtag: Agora o Arthur       volume of tweets: 34387
Hashtag: Riz Ahmed       volume of tweets: 27262
```

## 4.4 Data from Google trends

To get data from google we use a Python API called **pytrends.**

Unofficial API for Google Trends Allows simple interface for automating downloading of reports from Google Trend.

| | date | gaming | car | sports | actor | mobile | 5g | Covid-19 | Tesla | IPL | PESU | SRS | Jio | Nokia | Dell | FIFA | ICC | Fever | Mask | Lockdc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 05-04-20 | 57 | 57 | 47 | 76 | 84 | 100 | 45 | 28 | 4 | 57 | 38 | 56 | 37 | 89 | 39 | 26 | 100 | 31 | |
| 3 | 12-04-20 | 61 | 57 | 49 | 87 | 90 | 48 | 34 | 33 | 5 | 38 | 31 | 47 | 40 | 84 | 45 | 31 | 77 | 36 | |
| 4 | 19-04-20 | 62 | 58 | 54 | 74 | 90 | 30 | 30 | 35 | 4 | 18 | 40 | 100 | 38 | 87 | 37 | 25 | 76 | 37 | |
| 5 | 26-04-20 | 68 | 58 | 55 | 90 | 92 | 26 | 29 | 47 | 3 | 54 | 43 | 7 | 44 | 89 | 40 | 39 | 58 | 41 | |
| 6 | 03-05-20 | 65 | 61 | 53 | 84 | 93 | 24 | 27 | 42 | 3 | 18 | 41 | 73 | 42 | 90 | 38 | 28 | 56 | 36 | |
| 7 | 10-05-20 | 64 | 67 | 53 | 79 | 92 | 22 | 35 | 38 | 3 | 18 | 48 | 37 | 41 | 91 | 34 | 24 | 49 | 52 | |
| 8 | 17-05-20 | 66 | 74 | 56 | 82 | 94 | 19 | 30 | 34 | 5 | 55 | 56 | 80 | 43 | 88 | 30 | 22 | 53 | 33 | |
| 9 | 24-05-20 | 64 | 82 | 57 | 75 | 87 | 20 | 29 | 39 | 4 | 39 | 53 | 53 | 41 | 86 | 27 | 19 | 71 | 26 | |
| 10 | 31-05-20 | 60 | 82 | 57 | 73 | 93 | 17 | 32 | 45 | 4 | 0 | 38 | 41 | 43 | 94 | 26 | 24 | 73 | 41 | |
| 11 | 07-06-20 | 66 | 84 | 63 | 84 | 92 | 17 | 32 | 51 | 3 | 54 | 70 | 48 | 40 | 82 | 28 | 26 | 54 | 40 | |
| 12 | 14-06-20 | 61 | 84 | 95 | 76 | 94 | 16 | 33 | 44 | 3 | 36 | 42 | 68 | 44 | 90 | 29 | 20 | 53 | 36 | |

# 5) CHAPTER 5

## PROJECT REQUIREMENT SPECIFICATION

---

Forecasting is the phenomenon of predicting something based on its present and past data and commonly by analysis of trends.

In the process of forecasting we usually collect the raw data which are most trending in the present to make the future analysis so the efficient data collection method is required to collect the data along with the timestamp within it. So, we glean the data from various social media websites like twitter which provides the trends of the topics and google trends API which provides keywords trend with the topic week stamp and we can customise the location and finally, we extracted data from the YouTube which provide data of videos which has most views and its likes counts and comments. The data extracted completely consist of text data which has to be undergone for deep pre-processing.

In the preprocessing where the collected data are cleaned using various pre-processing techniques where the data is subjected to stemming which helps in reducing the word into its root form and stop word removal where words which are having no use in analysis, tokenization where words are converted into transactions, normalization is done to remove the noise from the words.

Then the pre-processed topics are classified into respective domains using text classification techniques like NLP techniques, SVM text classification, neural networks. In this classification where cars are classified into automobiles, Nokia is classified into mobiles and cricket is classified into sports domain.

In future selection where the collected data is optimised by selecting the subset of features to use in the model to forecast which also helps in minimising the overfitting problem.

Forecasting data here we forecast the prepared data set using the forecasting techniques like poisons process for the variation of the particular topics over a period of time, Markov chain for trend analysis, Apriori algorithms for frequent item set generation, then neural network for prediction.

Testing using RMSE score to evaluate the model accuracy. Interpreting the results using the visual graphs and trend score for each topic.

# 6) CHAPTER 6

## SYSTEM REQUIREMENTS SPECIFICATION

---

## Functional Requirements:

**1)Data Extraction:**

The system has to perform the extraction of the data from the various social media platform like twitter and YouTube and based on the keywords obtained from the fore mentions we see the trend of topic in google trends. Every mentioned thing is done by using the API provided by the individual platforms. Every data extracted is stored in the .csv file.

**2)Pre-processing:**

Then the system has to perform the pre-process of the extracted data which has to be done in real time to provide real time prediction results. Mainly pre-processing of data includes normalisation, removing stop words and small words, stemming and removing verbs.

**3)Text classification:**

Classification of topics into the respective domain they belong. This is done by using the text classification techniques such as NLP techniques, SVM classifiers and also by using neural network classifiers like by using RNN.

**4)Forecasting:**

The pre-processed and classified data then forecasted using the forecasting model which predicts the forthcoming trending topics. Here, we use the NN model for forecasting which gives us high accurate results.

**5)Visual representation of results:**

Then based on the prediction of the model system going to picturise the results in interactive graphs using matplotlib and word cloud.

**Performance Requirement:**

The dataset has to be generated and pre-processing of generated data, then classification of topics from the dataset and forecasting of the topics followed by displaying the results has to be done in real time. So, to perform all aforementioned tasks the system should have high performance to carry out all this process in background. The system also should be available 24x7 to the customers.

**Security Requirement:**

No confidential details are kept so there is much less need to worry about the security of the system but DOS are to be noticed.

**Safety Requirements:**

The dataset generated is backed up safely and maintained without any corruption of generated data set contents has to be taken care.

**Software Requirement**

We use a drive to store the extracted data and then the modules are used to pre-process the data. Initial testing done in either Jupyter notebook or Google Collab. Libraries are also used like Keras and Tensorflow.

**Hardware Requirements:**

The system contains GPU for better parallel processing and free space to hold the generated dataset.

7) CHAPTER 7

# SYSTEM DESIGN

---

### 7.1 Proposed methodology

**1. Data Extraction:**

Using the api provided by different social media platforms we glean the data from the twitter of trending topics, YouTube trends and google trends. The extracted data is stored in a .csv file for further analysis.

**2. Pre-processing:**

The data is pre-processed using the various pre-processing techniques.

1) First the normalization of data is done to remove the noise from the dataset.

2)After stemming is done to make the topic reduced into the root form.

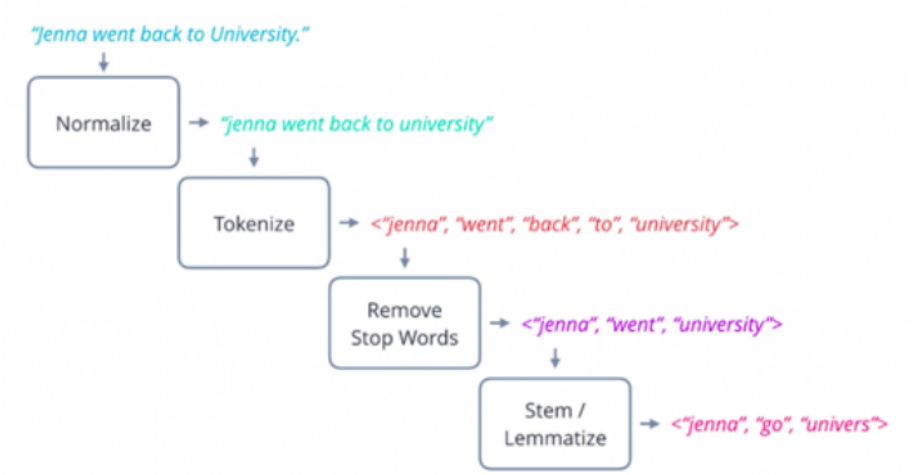
3)Removal of stop words from dataset.

4)Tokenization

5)Part of speech tagger

6)Removing verbs.

7)Lemmatization.




**3. Topic classification:**

Then the pre-processed data is subjected for the classification where topics are classified to respective domains. In this classification where cars are classified into automobiles, Nokia is classified into mobiles and cricket is classified into sports domain.


**4. Future Selection:**

In future selection where the collected data is optimised by selecting the subset of features to use in the model to forecast which also helps in minimising the overfitting problem.

**5. Forecasting topics:**

Forecasting data here we forecast the prepared data set using the forecasting

techniques:

Poisons process for the variation of the particular topics over a period of time,

Markov chain for trend analysis.

Apriori algorithms for frequent item set generation.

Neural network for prediction.

Clustering for violations detection.

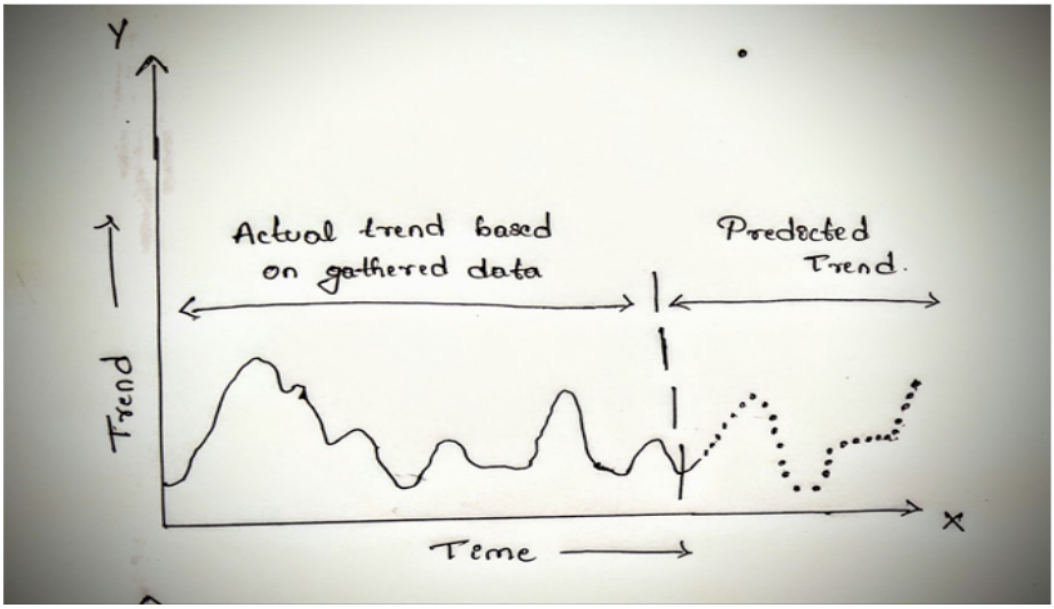
**6. Evaluating Model Accuracy:**

Testing using RMSE score to evaluate the model accuracy.RMSE is the standard deviation of errors or residuals.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - F_t)^2}$$

**7. Results:**

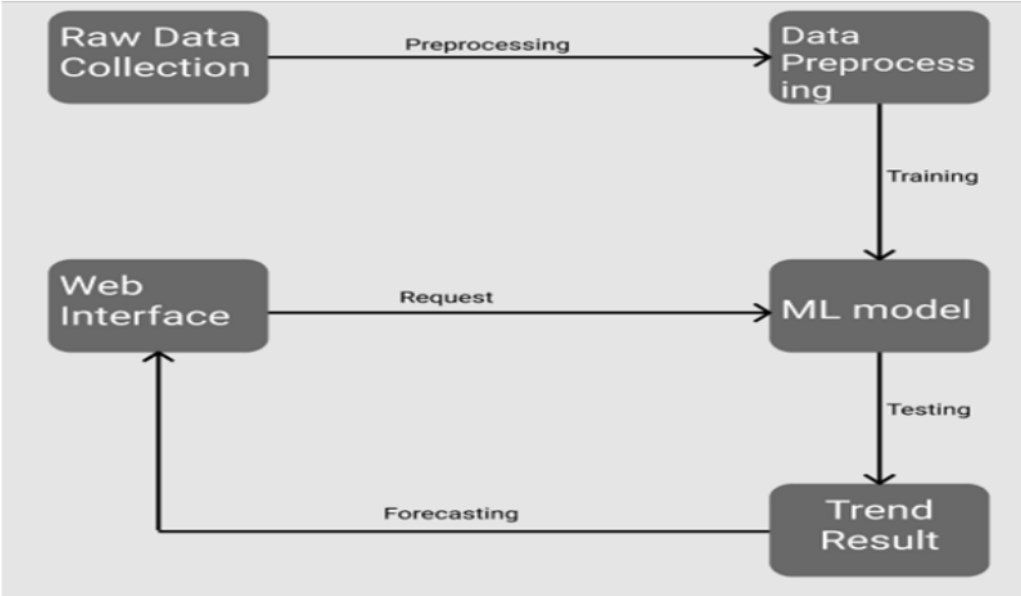
Interpreting the results using the visual graphs and trend score for each topic. We are using matplotlib, seaborn and word cloud models to graphically present the results. The interpretation of data using all results finally done.




**8)Microservices:**

The Flask API user interface is developed for clients. Microservices which connect every part of the system for exchange of data.

FINAL SYSTEM DESIGN



# 8) CHAPTER 8

## IMPLEMENTATION AND PSEUDOCODE

Until now we have done with our extraction of data from twitter, google trends and youtube using specific API provided by the platforms.

Our Pseudocode for this entire project is:

1.Extract data from social networks using web scraping and api's.

2.Preprocess Data using NLP and Normalisation ,Removal of POS, stop words, stemming etc.

3.Using extracted Data build a model for training

4.Train the model.

5.Test the trained Model and update the model as required.

### Extracting data from Youtube API

```
import requests, sys, time, os, argparse


# List of simple to collect features
snippet_features = ["title",
                    "publishedAt",
                    "channelId",
                    "channelTitle",
                    "categoryId"]


# Any characters to exclude, generally these are things that become problematic in CSV files
```

```
unsafe_characters = ['\n', '"']

# Used to identify columns, currently hardcoded order
header = ["video_id"] + snippet_features + ["trending_date", "tags", "view_count", "likes", "dislikes",
                                            "comment_count", "thumbnail_link", "comments_disabled",
                                            "ratings_disabled", "description"]



def setup(api_path, code_path):
    with open(api_path, 'r') as file:
        api_key = file.readline()

    with open(code_path) as file:
        country_codes = [x.rstrip() for x in file]

    return api_key, country_codes



def prepare_feature(feature):
    # Removes any character from the unsafe characters list and surrounds the whole item in quotes
    for ch in unsafe_characters:
        feature = str(feature).replace(ch, "")
    return f'"{feature}"'



def api_request(page_token, country_code):
    # Builds the URL and requests the JSON from it
    request_url =
f"https://www.googleapis.com/youtube/v3/videos?part=id,statistics,snippet{page_token}chart=mostPo
pular&regionCode={country_code}&maxResults=50&key=AIzaSyB5zq59mf-ZcG3r9MbbkdWYjEX8e6
Cwnak"
    request = requests.get(request_url)
```

```
    if request.status_code == 429:
        print("Temp-Banned due to excess requests, please wait and continue later")
        sys.exit()
    return request.json()
def get_tags(tags_list):
    # Takes a list of tags, prepares each tag and joins them into a string by the pipe character
    return prepare_feature("|".join(tags_list))
def get_videos(items):
    lines = []
    for video in items:
        comments_disabled = False
        ratings_disabled = False


        # We can assume something is wrong with the video if it has no statistics, often this means it has
been deleted
        # so we can just skip it
        if "statistics" not in video:
            continue


        # A full explanation of all of these features can be found on the GitHub page for this project
        video_id = prepare_feature(video['id'])


        # Snippet and statistics are sub-dicts of video, containing the most useful info
        snippet = video['snippet']
        statistics = video['statistics']


        # This list contains all of the features in snippet that are 1 deep and require no special
processing
        features = [prepare_feature(snippet.get(feature, "")) for feature in snippet_features]


        # The following are special case features which require unique processing, or are not within the
snippet dict
```

```
        description = snippet.get("description", "")
        thumbnail_link = snippet.get("thumbnails", dict()).get("default", dict()).get("url", "")
        trending_date = time.strftime("%y.%d.%m")
        tags = get_tags(snippet.get("tags", ["[none]"]))
        view_count = statistics.get("viewCount", 0)


        # This may be unclear, essentially the way the API works is that if a video has comments or
ratings disabled
        # then it has no feature for it, thus if they don't exist in the statistics dict we know they are
disabled
        if 'likeCount' in statistics and 'dislikeCount' in statistics:
            likes = statistics['likeCount']
            dislikes = statistics['dislikeCount']
        else:
            ratings_disabled = True
            likes = 0
            dislikes = 0


        if 'commentCount' in statistics:
            comment_count = statistics['commentCount']
        else:
            comments_disabled = True
            comment_count = 0


        # Compiles all of the various bits of info into one consistently formatted line
        line = [video_id] + features + [prepare_feature(x) for x in [trending_date, tags, view_count, likes,
dislikes,
                                                                     comment_count, thumbnail_link, comments_disabled,
                                                                     ratings_disabled, description]]
        lines.append(",".join(line))
    return lines
```

```
def get_pages(country_code, next_page_token="&"):
    country_data = []

    # Because the API uses page tokens (which are literally just the same function of numbers everywhere) it is much
    # more inconvenient to iterate over pages, but that is what is done here.
    while next_page_token is not None:
        # A page of data i.e. a list of videos and all needed data
        video_data_page = api_request(next_page_token, country_code)
        if "error" in  video_data_page.keys():
            print("Error retrieving data")
            print(video_data_page)
            return
        # Get the next page token and build a string which can be injected into the request with it, unless it's None,
        # then let the whole thing be None so that the loop ends after this cycle
        next_page_token = video_data_page.get("nextPageToken", None)
        next_page_token = f"&pageToken={next_page_token}&" if next_page_token is not None else next_page_token

        # Get all of the items as a list and let get_videos return the needed features
        items = video_data_page.get('items', [])
        country_data += get_videos(items)

    return country_data


def write_to_file(country_code, country_data):

    print(f"Writing {country_code} data to file...")

    6
    if not os.path.exists(output_dir):
```

```
        os.makedirs(output_dir)


    with open(f"{output_dir}/{time.strftime('%y.%d.%m')}_{country_code}_videos.csv", "w+",
encoding='utf-8') as file:
        for row in country_data:
            file.write(f"{row}\n")



def get_data():
    for country_code in country_codes:
        data = get_pages(country_code)
        if data is not None:
            country_data = [",".join(header)] + data
            write_to_file(country_code, country_data)
        else:
            print("fatal error")
if __name__ == "__main__":


    parser = argparse.ArgumentParser()
    parser.add_argument('--key_path', help='Path to the file containing the api key, by default will use
api_key.txt in the same directory', default='api_key.txt')
    parser.add_argument('--country_code_path', help='Path to the file containing the list of country
codes to scrape, by default will use country_codes.txt in the same directory',
default='country_codes.txt')
    parser.add_argument('--output_dir', help='Path to save the outputted files in', default='output/')


    args = parser.parse_args()


    output_dir = args.output_dir
    api_key, country_codes = setup(args.key_path, args.country_code_path)


    get_data()
```

# 9) CHAPTER 9

## CONCLUSION OF CAPSTONE PROJECT PHASE-1

---

Collection of research paper in this subject- Jan/21

Analysis of Collected research paper along with team-Feb/21

Formalisation of extraction data using API and Finding API's Feb/21

Data Extraction using API and Discussion on model for forecasting -Mar/21

Discussion on interpreting results and microservices - Apr/21

Discussion on what additional things can be added -Apr/21

Preparing to make project to tackle business problems -Apr/21

## 10) Chapter 10

### PLAN OF WORK FOR CAPSTONE PROJECT Phase -2

Finding the method to organize all the extracted data.

Extensive Preprocessing of the collected Data.

Extraction and Classification of the topics from data.

Feature selection from the data.

Building models for forecasting of text data.

Training and Evaluating model.

Building a user interface.

Building microservices and deploying the models.

# BIBLIOGRAPHY


1. Jose L. Hurtado*† , Ankur Agarwal† and Xingquan Zhu†,

”Topic discovery and future trend forecasting for texts”

2. Roselina Sallehuddin, Siti Mariyam Hj. Shamsuddin, Siti Zaiton Mohd. Hashim,,Ajith Abrahamy, “Forecasting Time Series Data Using Hybrid GREY Neural Network And ARIMA Model”.

3. “YouTube View Prediction with Machine Learning”
Int. J. Business Information Systems, Vol. 13, No. 3, 2013 359

4. Conrad Tucker1 and Harrison M. Kim1 (1)
“PREDICTING EMERGING PRODUCT DESIGN TREND BY MINING PUBLICLY AVAILABLE CUSTOMER REVIEW DATA”.- University of Illinois at Urbana-Champaign, USA.

# Thank You

# “Feature trend forecasting using text data”

ORIGINALITY REPORT

| 6% | 4% | 3% | 1% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Pinto, Henrique, Jussara M. Almeida, and Marcos A. Gonçalves. "Using early view patterns to predict the popularity of youtube videos", Proceedings of the sixth ACM international conference on Web search and data mining - WSDM 13 WSDM 13, 2013.<br>Publication | 2% |
| 2 | dli.iiit.ac.in<br>Internet Source | 1% |
| 3 | www.coursehero.com<br>Internet Source | 1% |
| 4 | eprints.utm.my<br>Internet Source | <1% |
| 5 | towardsdatascience.com<br>Internet Source | <1% |
| 6 | learn.pyblish.com<br>Internet Source | <1% |
| 7 | Submitted to Rochester Institute of Technology<br>Student Paper | <1% |

| | | |
|---|---|---|
| 8 | github.com<br>Internet Source | <1% |
| 9 | www.ijbiotech.com<br>Internet Source | <1% |
| 10 | Roselina Sallehuddin, Siti Mariyam Hj. Shamsuddin. "HYBRID GREY RELATIONAL ARTIFICIAL NEURAL NETWORK AND AUTO REGRESSIVE INTEGRATED MOVING AVERAGE MODEL FOR FORECASTING TIME-SERIES DATA", Applied Artificial Intelligence, 2009<br>Publication | <1% |
| 11 | journalofbigdata.springeropen.com<br>Internet Source | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | < 5 words |
| Exclude bibliography | On | | |