HIGH LEVEL DESIGN DOCUMENT

Forecasting future trending topics using text data.

UE18CS390A – Capstone Project Phase – 1

Submitted by:

Kavishankar K S	**PES1201802001**
Mohammed Zeeshan	**PES1201801814**
Rohan M	**PES1201801798**

Under the guidance of

Prof. Dinesh Singh
Designation
PES University

January - May 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

TABLE OF CONTENTS

Note:

Section – 1 & Section 2	**Common for Product Based and Research Projects**
Section 3 to Section 11	**High-Level Design for Product Based Projects.**
Section 12	**High-Level Design for Research Projects.**
Appendix	**Provide details appropriately**

1. Introduction

The growth in popularity of web 2.0 applications 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 amount of content uploaded on YouTube in 60 days is equal to the amount of data broadcasted for 60 years by ABC , NBC, CBS together . This shows the content upload rate, This shows that popularity across different contents is undistributed on Youtube. It also has been reported that youtube users upload almost 72hours of video content every minute. This shows how important it is to forecast future trends.

In this condition , Forecasting content popularity will be of greater significance to support and drive various design and management of services. For example, 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.

Predicting Web content popularity is an important task for supporting the design and evaluation of a wide range of systems, from targeted advertising to active search and recommendation services. We here present two simple models for predicting the future popularity of Web content based on historical information given by early popularity measures.

2. Current System [if applicable]

No correct related work has been done in this field.

3. Design Considerations

3.1. Design Goals

System is to be designed in a fully automated way where there is less maintainability and Should be very high to interpret the results in real time.

3.2. Architecture Choices

- We propose to use an Iterative methodology to solve this problem. This approach follows all the steps of a waterfall model but in repetitive cycles called 'iterations'.
- Unlike agile, the iterative model requires less customer involvement and has a pre-defined scope of increments. Yet it still shares the same goals as an agile model.
- It produces a working version of the software early on the process and accommodates changes between increments.The final product is developed by working on smaller chunks during each iteration.
- Generates working software quickly and early during the software life cycle.
- Easier to test and debug during a smaller iteration.
- Easier to manage risk because risky pieces are identified and handled during its iteration.
- ***Alternate choices:*** *Waterfall model* approach

Cons of using the waterfall model over agile model

- No working software is produced until late during the life cycle.
- High amounts of risk and uncertainty.
- Not a good model for complex and object-oriented projects.
- Poor model for long and ongoing projects.
- Not suitable for the projects where requirements are at a moderate to high risk of changing. So, risk and uncertainty is high with this process model.
- It is difficult to measure progress within stages.
- Cannot accommodate changing requirements.
- Adjusting scope during the life cycle can end a project.
- Integration is done as a "big-bang. at the very end, which doesn't allow identifying any technological or business bottleneck or challenges early.

3.3. Constraints, Assumptions and Dependencies

Constraints:

- As our project is based on real time data.For every result preprocessing and extraction of data is required.
- Since we are using ML models to predict the future trends, we may need huge time and high computational power if we use large datasets to train the model.
- A lot of time is consumed if we want to make changes to the model depending on the circumstances, since training the model requires a large amount of time.
- We are dealing with text data which is indeed a big constraint since we can't work on it, so we have to convert into some form before any process.
- Extent of forecasting.

Assumptions:

- The trending topics in future are based on the current and past trend.
- The frequency of occurrence of topics helps in predicting it's trend.
- Accuracy of the chosen model is high.
- Topics are low resolution and high sensitive.
- Rising and falling of the topic do not have a long term span.

Dependency:

- Confounding variables in data.

  Effects the both dependent and independent variable.

- Correlation between topics.

  Extent of relation between each topic.

- Data Quality .

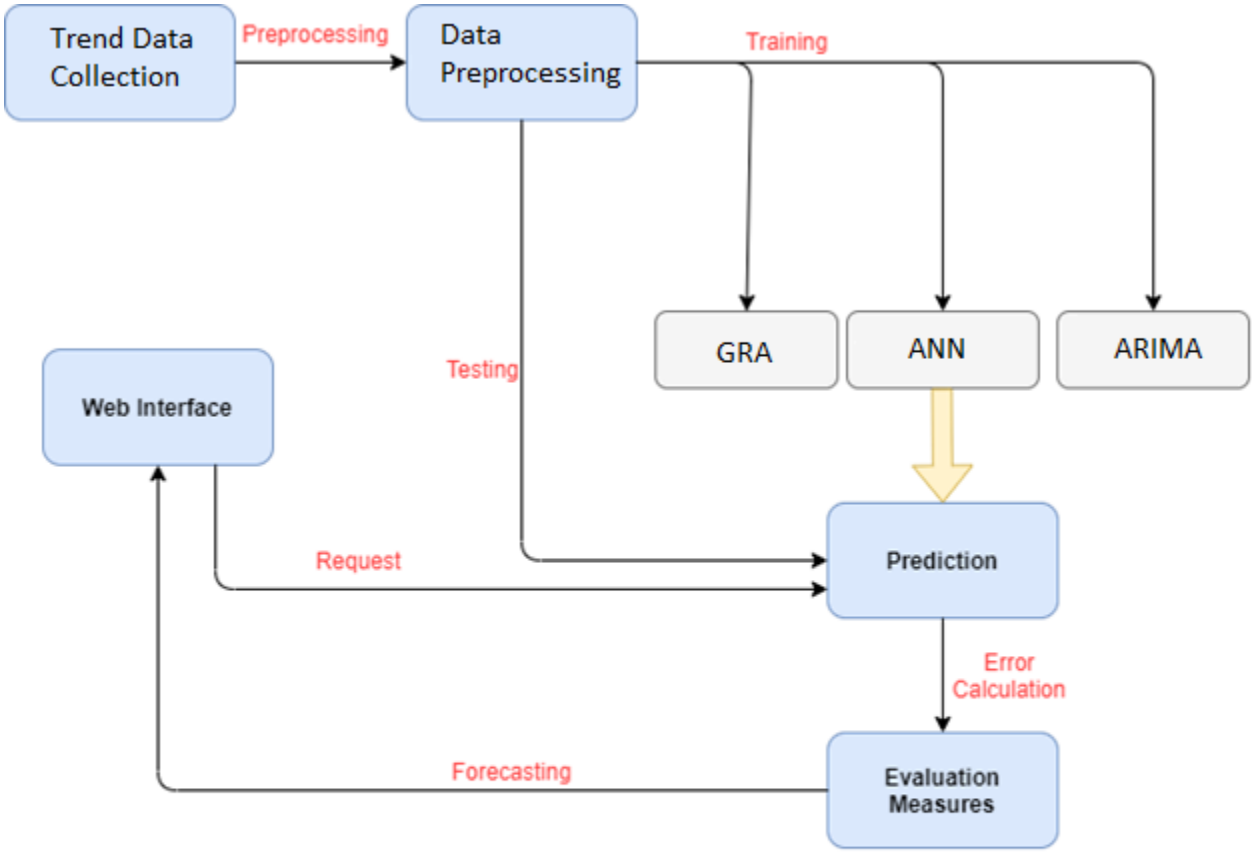
  Large dataset and useful features in the dataset.

- Temporal dependency.

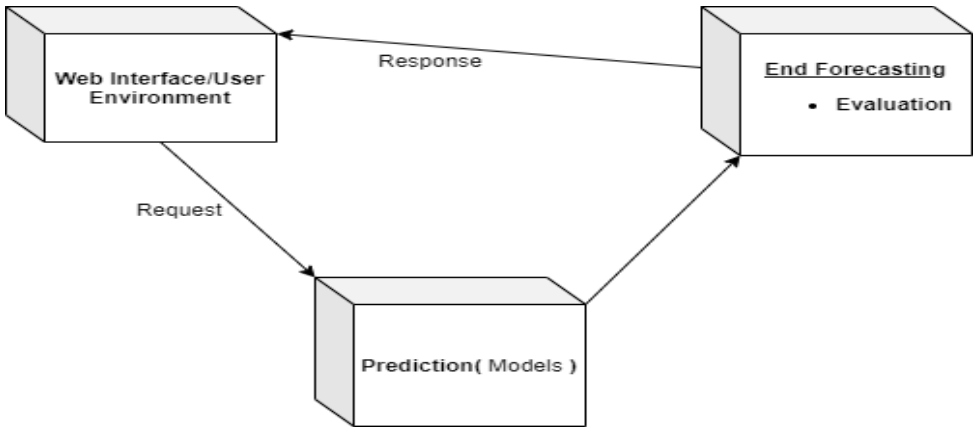
  Casual relationship b\w the topics.

4. High Level System Design

TENTATIVE HLD:

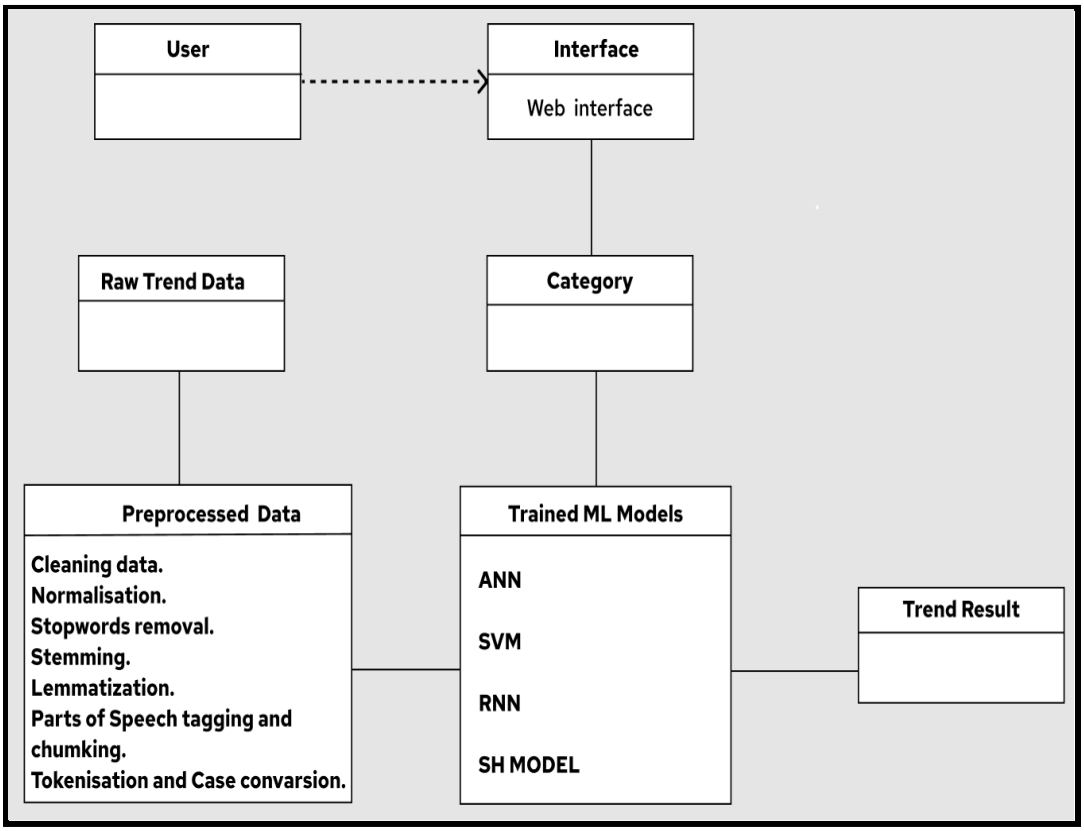


Deployment diagram:



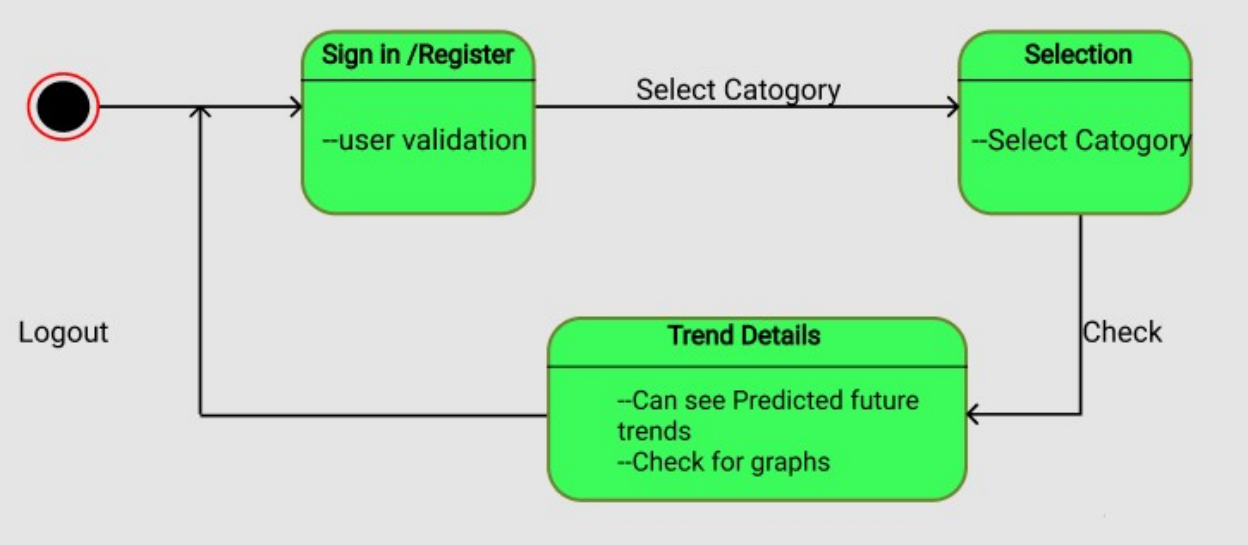
5. Design Description

5.1. Master Class Diagram

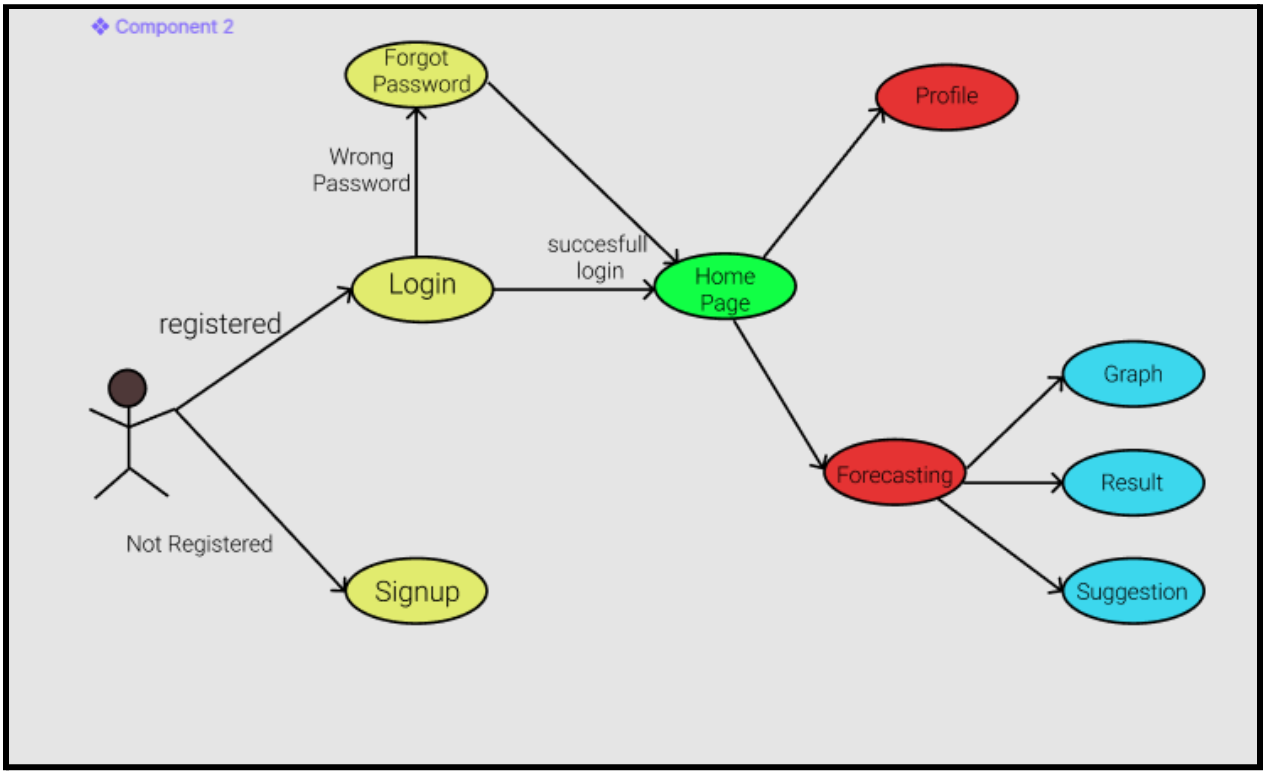


6. ER Diagram / Swimlane Diagram / State Diagram (include as appropriate)

State diagram:

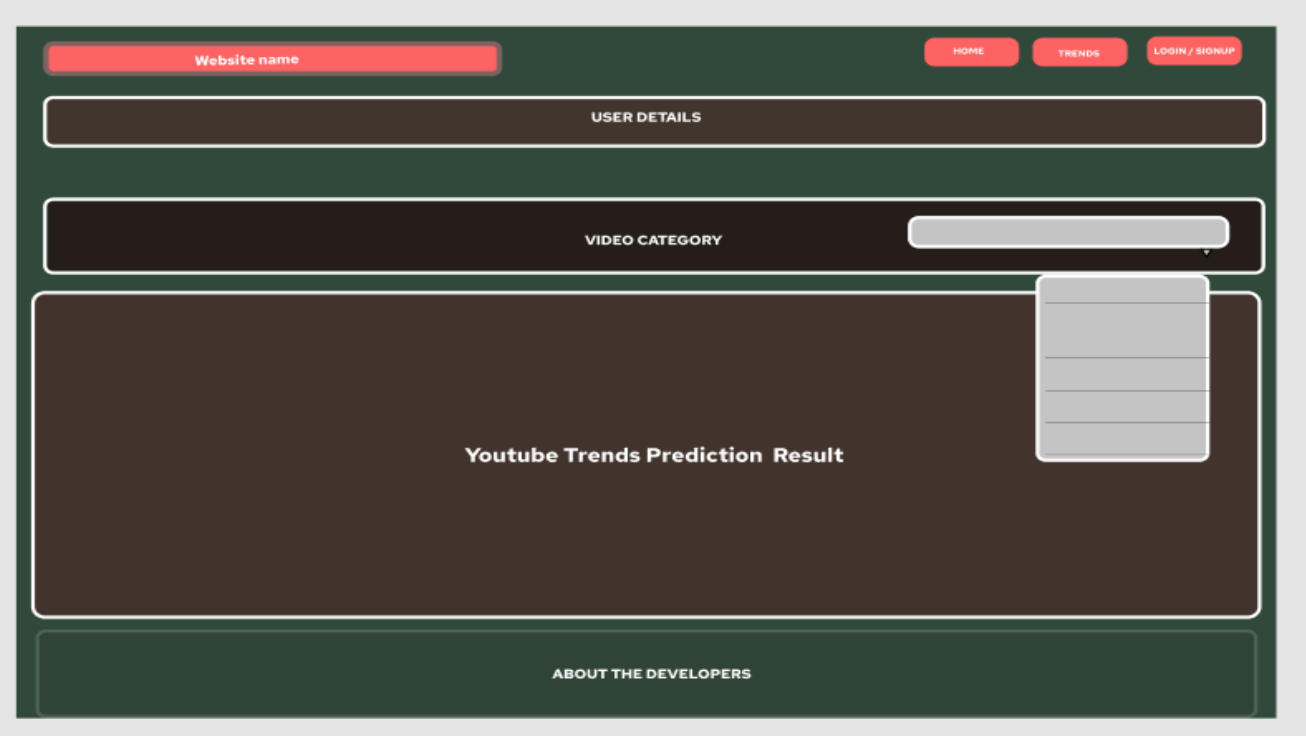


Use case diagram:

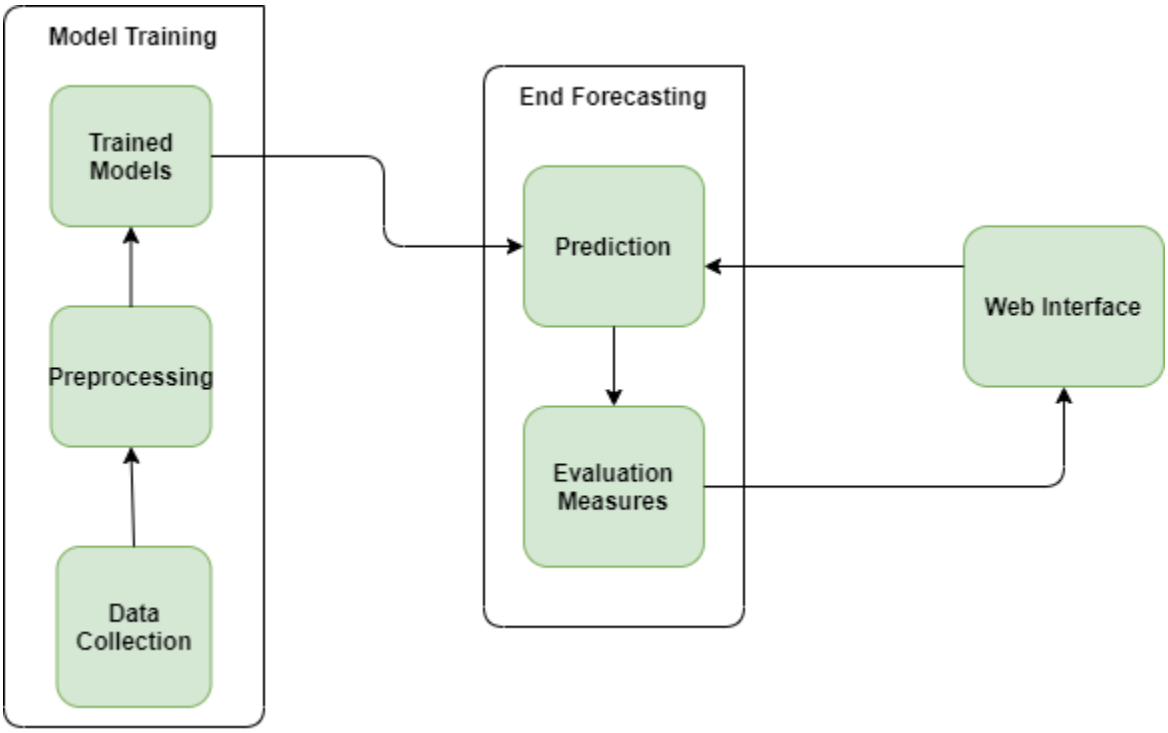


#	Entity	Name	Definition	Type
ENTITIES				
1.	USER	user_id		string
#	**Attribute**	**Name**	**Definition**	**Type (size)**
DATA ELEMENTS				
1.	user_name		user name	string(10)
2.	user_id		id of user	string(10)
3.	user_email		email of user	string(10)
4.	user_pass		password of user	string(10)

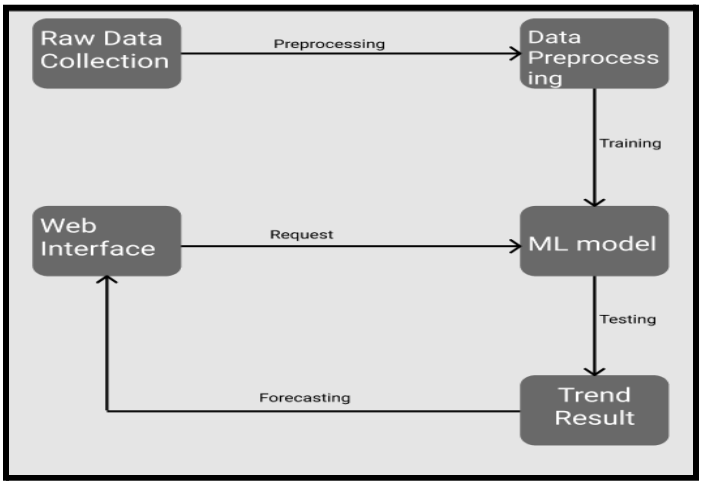
7. User Interface Diagrams



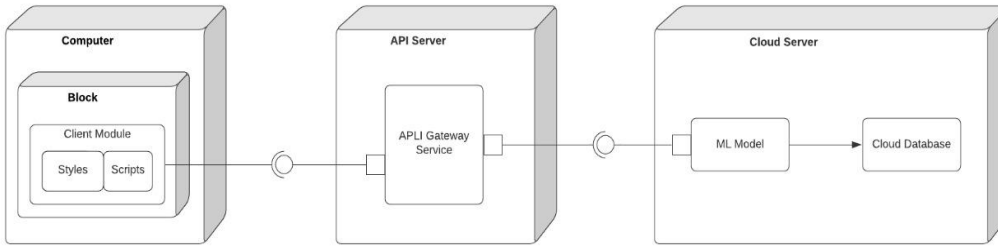
8. External Interfaces



9. Packaging and Deployment Diagram



DEPLOYMENT DIAGRAM



Deployment diagram 2.0

10. Help

Documentation of the project will be available where all the context sensitive help will be provided. Detailed working of the system will be there in the documentation. Further support can be given by approaching the developers

The user support is provided for customers to access all the results and trends.

11. Design Details

1.1. Novelty

We introduced the new way of accessing and gathering data from the social media websites using the API. We aim to build a good hybrid machine learning model for prediction.

1.2. Innovativeness

Providing variation of trend of a particular topic over a period of time.

1.3. Interoperability

The system is designed in a way that uses the dataset stored in the cloud and by processing the data produce the results.

1.4. Performance

We hope to think that the accuracy produced by the chosen model is high enough to produce good results and using GPU we can enhance the performance which helps parallel computing**.**

1.5. Security

Security is not a very high concern as of now.

1.6. Reliability

We consider the fact that the system is highly reliable as there is no loss of performance and information also in security.

1.7. Maintainability

Very less maintenance is needed because we are going to automate everything including data extraction and data preprocessing and forecasting.

1.8. Portability

As it is a web application this can be accessible from the modern browser.

1.9. Legacy to modernization:

Its highly valuable wealth in the field of decision making, forecasting demands and precautions can be taken based on trend.

1.10. Reusability

The model used for forecasting, gathered dataset and preprocessing application can be reusable to build the system as aforementioned things.

Appendix A: Definitions, Acronyms and Abbreviations

GRA-Grey relational analysis
ANN-Artificial neural network
ARIMA-Auto regressive integrated moving average

Appendix B: References

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.

Appendix D: Traceability Matrix

[Demonstrate the forward and backward traceability of the system to the functional and non-functional requirements documented in the Requirements Document.]

Project Requirement Specification Reference Section No. and Name.	DESIGN / HLD Reference Section No. and Name.