PROJECT REQUIREMENT SPECIFICATION

Forecasting is where phenomena 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 way is required to collected the data along with the timestamp within it. So, we glean the data from varies 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 from completely consist of text data which has to be undergone for deep pre-processing.

In the pre-processing where the collected data are cleaned using varies 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 domain 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 feature 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.

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 in stored in the .csv file.

2)Pre-processing:

Then 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 include normalisation, removing stop words and small words, stemming and removing verbs.

3)Text classification:

Classification of topics into 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 classifier 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 NN model for forecasting which gives us high accurate results.

5)Visual representation of results:

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

Performance Requirement:

The dataset which has to be generated and pre-processing of generated data, then classification of topics from dataset and forecasting of the topics followed 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 available 24x7 to the customers.

Security Requirement:

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

Safety Requirements:

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

Software Requirement

We use 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 containing GPU for better parallel processing and free space to hold the generated dataset.

SYSTEM DESIGN

1. Data Extraction:

   Using the api provided different social media platform we glean the data from the twitter of trending topics, YouTube trends and google trends. The extracted data is in stored in .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.

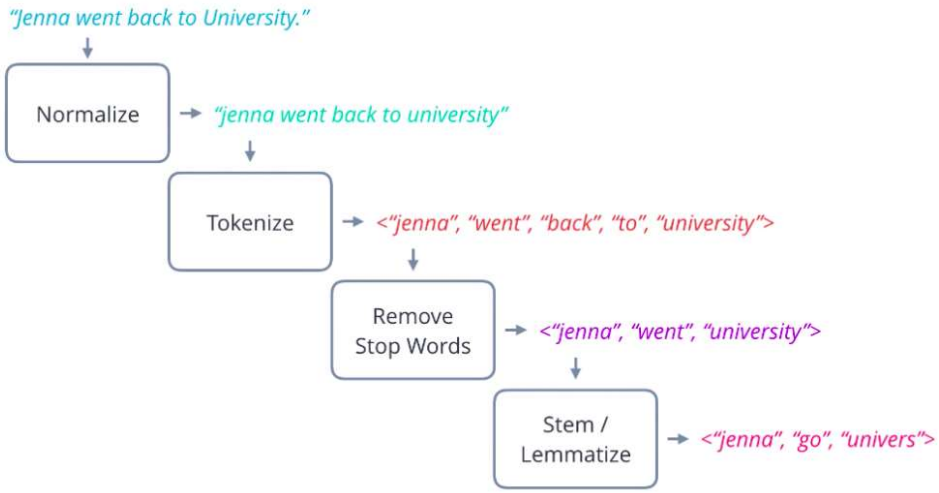
   3)Removal of stop words from dataset.

   4)Tokenization

   5)Part of speech taggers

   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 feature 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:

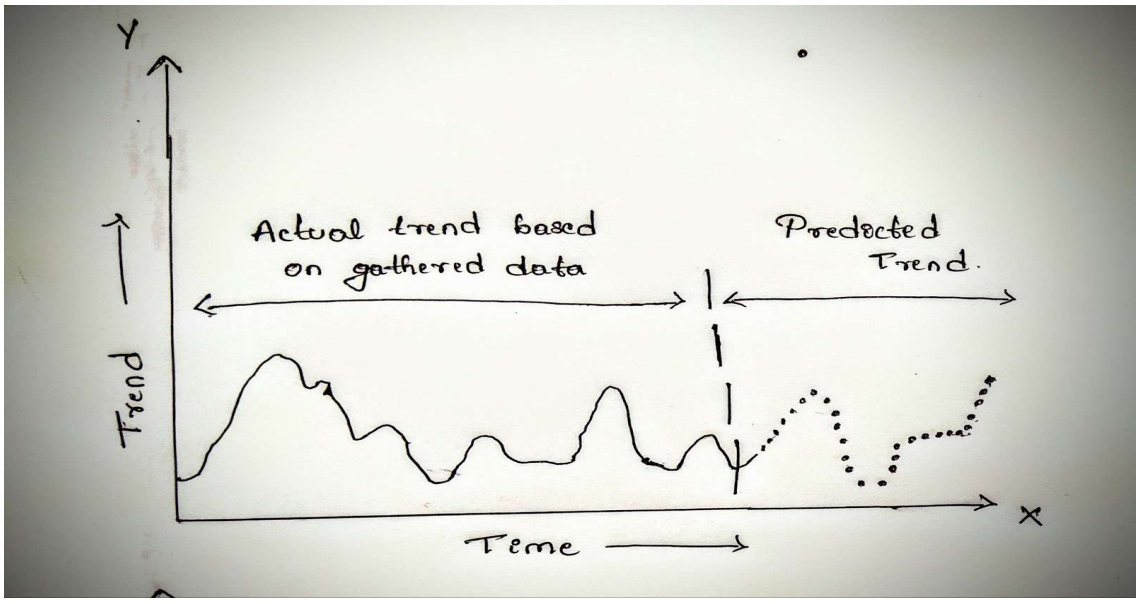
   Testing using RMSE score to evaluate the model accuracy.

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

RMSE is the standard deviation of errors or residuals.

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 graphical present the results. The interpretation of data using all results finally done.



8)Microservices:

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

FINAL SYSTEM DESIGN