Facebook status prediction

Project Report.

s.nitish raj

Objective:

Development of a predictive model for monitoring the people who are going through difficulties in their adolescence from their Facebook status The model will determine what is the emotional state of the people through their Facebook post

Benefits:

1. Detection of upcoming suicide attempts

2. Gives better insight into the emotional state of the children to their parents

3. we can identify any crime is committing by people who are of adolescent age.

Data Sharing Agreement :

Sample file name (ex emotion.csv)

Length of date stamp(8 digits)

Length of time stamp(6 digits)

Number of Columns

Column names

Column data type

Architecture

1. start	2. data for Facebook status	3. web scrapping	4. data transformation	5. data insertion in databases
6. export data from the database	7. data preprocessing	8.nlp techniques	9. data vectorization	10. model building
11. cloud setup	12. pushing app to the cloud	13. application start	14. data from user	15. data validation
16. data insertion into the database	17. bring data into prediction	18. model call from a prediction	19. Facebook status prediction	20. saving output at the database

Data Validation and Data Transformation :

data validation and data transformation play important roles in ensuring the accuracy and reliability of the predictive model.

Data validation involves checking the quality and completeness of the data that will be used to train the predictive model. This includes identifying missing or incorrect data, as well as ensuring that the data is consistent and follows the expected format. In the context of Facebook status prediction, this may involve checking that the status updates are written in a consistent language and format, that there are no missing values in the data, and that there are no duplicates or errors.

Data transformation, on the other hand, involves preparing the data for use in the predictive model. This may involve cleaning and pre-processing the data, as well as transforming it into a format that is suitable for the model. In the context of Facebook status prediction, this may involve converting the text of the status updates into a numerical representation that can be used by the model, such as a count vectorization or word embedding representation.

Overall, data validation and data transformation are important steps in the data preparation process for Facebook status prediction. By ensuring that the data is of high quality and in a suitable format, we can improve the accuracy and reliability of the predictive model.

Data Insertion in Database:

Table creation:- Table name “emotion_status" is created in the database for

inserting the files. If the table is already present then new files are inserted in the

same table.

Insertion of files in the table - All the files in the "facebook_data_folder" are inserted

in the above-created table. If any file has an invalid data type in any of the columns,

the file is not loaded in the table

Model Training:

Data Export from Db :

1.The accumulated data from db is exported in CSV format for model training

Data Preprocessing

2. Performing EDA to get an insight of data like identifying distribution, outliers, trends

among data etc.

3. Check for null values in the columns. If present impute the null values.

4. Encode the categorical values with numeric values.

5. Perform Standard Scalar to scale down the values.

Logistic Regression:

Logistic regression is a statistical model that can be used to predict the probability of a binary outcome based on one or more predictor variables. In the context of Facebook status emotion prediction, logistic regression can be used to model the relationship between various features of a status update, such as the language used, the presence of certain keywords, the length of the post, and the probability of the post eliciting a specific emotional response, such as happiness, sadness, or anger.

To use logistic regression for Facebook status emotion prediction, we first need to collect and prepare a dataset of status updates and their associated emotional responses. We can then select a set of features that we believe may be predictive of the emotional response and use logistic regression to estimate the coefficients for each feature.

Once we have trained the logistic regression model, we can use it to make predictions on new status updates. We simply input the values of the selected features for the new status update into the model, and it will output a probability of the post eliciting a specific emotional response. We can then use a threshold value to convert this probability into a binary prediction of whether the post will elicit an emotional response or not.

Overall, logistic regression is a useful tool for predicting emotional responses to Facebook status updates. By modeling the relationship between various features of the post and the probability of eliciting a specific emotional response, we can gain insights into the factors that influence emotional reactions on the platform and improve our understanding.

Model Selection :

after applying some of the machine learning algorithms used for text classification like decision trees, random forest, naive bias, and logistic regression. the highest score is achieved by the logistic regression so i have selected logistic regression and the model also generalized well.

Prediction:

1, The testing files are shared in the batches and we perform the same Validation operations, data transformation, and data insertion on them.

2, The accumulated data from db is exported in CSV format for prediction

3. We perform data pre-processing techniques on it.

.4.logistic regression model created during training is loaded and the preprocessed data is predicted

5. Once the prediction is done for all the status text. The predictions are saved in CSV format and shared.

Q & A:

Q1) What's the source of data?

The data for training is provided by the client in multiple batches and each batch contain multiple files

Q 2) What was the type of data?

The data was a combination of text and Categorical values.

Q 3) What's the complete flow you followed in this Project?

Refer to slide model architecture in this report for a better Understanding

Q 4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q 5) How logs are managed?

We are using different logs as per the steps that we follow in validation and modeling like File validation log, Data Insertion, Model Training log, prediction

log etc.

Q 6) What techniques were you using for data pre-processing?

1,removed stopwords removed punctuation, remove special characters,null values, tokenization, count vectorization.tf-idf and label encoder

2, Removing outliers

3, Cleaning data and imputing if null values are present.

4. Converting categorical data into numeric values.

5. Scaling the data

Q 7) How training was done or what models were used?

1.after dividing data into test data and train data

2.The scaling was performed over training and validation data

3.Algorithms like the random forest, decision tree, naive bias, and logistic regression were used based on the score final model used for each and we saved that model.

Q 8) How Prediction was the done?

The testing files are shared by the client. We Perform the same life cycle till the data is cleaned. Then on the basis of the cleaned data model is loaded and performed prediction. n the end we get the accumulated data of predictions.