

Detail Project Report

Sentiment analysis

Dinesh Kumar Naik | Feedback Analyzer | 02-Aug-2021

# About the Project: -

Indian Oil is an Indian public sector oil and gas company. Project ePIC (e-Platform for Indian Oil Customers) is a project which attempts to establish a digital platform for devising and executing customer relationships and distributor/dealer management strategies and processes which leverage latest technological innovations enabling us to engage with our customers in better ways.

## Problem Statement:

Indian oil serves various emergency services (Oil, Petroleum Gas, Lubes) to the customer either direct or indirect channel. Due to the Govt. owned public company, customer satisfaction is high priority to improve the service. Now a day’s various digital channels available to the share there feedback, So company wanted to analysis the feedback and take the appropriate action on the same.

We use different unsupervised learning algorithms that can detect the topics in which all the comments will be grouped. The type of algorithms we use in the **modeling of topics**

## Value of the Project:

Value of the feedback is reflection of the service provided by the company, and according to the customer feedback company can change their strategy and modify the policy, so they can improve their service. But it is very difficult to read all the feedback manually, So the new AI Service (Modeling of topics) read all the feedback and come with the crisp high tending suggestions / complaints. Exp. In hilly area we have allow the domestic customer to take 4 cylinder refill, Launched tatkal (2 hours) cylinder refill seva for domestic customer.

This technique can be applied to any type of qualitative information.

## System Architecture:

Topic Modelling

DB

Pre-processing

Aspect Extraction

OBIEE Dashboard

Sentiment Analysis

Topic aspect Mapping

Above is the system architecture for the feedback analyzer.

**Storage:** Oracle 12g.

**Cloud: IBM Watson Cloud:**

**API:** Tweeter, IOCL Portal, Mobile App. SDMS CRM.

**Core tech Stack:** Topic modeling is as unsupervised machine learning technique that’s capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expression that best characterize a set of documents.

LDA (Latent Dirichlet Allocation) is an example of topic model and is used to classify text in a document to particular topic. It builds a topic per document model and word per topic model, modeled as Dirichlet distribution.

* **Data Pre-Processing:**
  + **Tokenization:** Split the text in to sentence and sentences into words. Lowercase the words and remove punction.
  + Words that have fewer than 3 characters are removed.
  + All **stopwords** are removed
  + Words are **lemmatized** – words in third person are changed to first person and verbs in past and future tenses are changed into present.
  + Words are **stemmed** – words are reduced to their root form.
* **Algorithms:**
  + BOW (Bag of Words)
  + TF-IDF
  + LDA (Latent Dirichlet Allocation)

**Hosting:**

**Dashboarding:** OBIEE

**Monitoring:**

**Hypercare:**

**Question Answers:**

**Q1. Tell me about your current project.**

The project is called classification of customer feedback review.

Objective of this project to listen the voice of customer via reviews, social media post, emails, chats and survey and responding in a way that will make them want to use our service or product again and even recommend it to others.

Our AI Model takes the review of the customer and label/classify the review against the each review. And based on that labeling, we do analysis the business positive / negative area as long as the customer demand.

**Q2. What was the size of the data.**

We do receive around 3-4 K review on daily basis.

**Q3. What was the data type.**

Data used for training is string and the string is converted into metrics to calculate the topic score.

**Q4. What was the team size and distribution.**

* 1 Solution Architect
* 1 DevOps engineers
* 1 QA Engineers
* 2 data scientists

**Q5. What is LDA (Latent Dirichlet Allocation).**

LDA is generative probabilistic model for collections of discrete data such as collection of written text.

Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

**Q6. What was the challenge you face during implementation of this model, and how you mitigate that challenge?**

End user review post the comments from different channel like portal, tweeter, mobile app. Some of the user write there issue in hindi font as well as Hindi sentence in English. They do mistake on spellings.

We tried our best to resolve the challenge implementing as third-party language conversion and auto spelling correction API.

**Q7. Where you have used Hypothesis Testing in your Machine learning Solution.**

Null hypothesis is a type of hypothesis used in statistics, where we treat everything same or equal.

During the experiment of the model performance, we use hypothesis testing to test whether result fall under the null hypothesis or alternat hypothesis.

**. Q8. What kind of statistical tests you have performed in your ML Application**

Below are statistical tests used in ML Application

* Hypothesis Testing
* P-Value Testing
* Z-Testing
* T-Testing
* Chi-Square Testing.

**Q9. What do you understand by P Value? And what is use of it in ML?**

P Value is the probability value based on this we test the null hypothesis to be true / false.

**Q10. Which type of error is severe Error, Type 1 or Type 2? And why with example.**

Type 1 error (False Positive) is the error, where we predict Positive but it is false and Type 2 error (False Negative) is the error, where we predict Negative but actually it is False.

* + - Regarding the severity, it depends on the case. When we analysis the human related case, Type 2 error is the severe, Ex. Diagnose the cancer case.
    - But when we analysis capital related case, Type 1 error is severe. Ex. Stock Market crash or false fire alarm.

**Q11. Where we can use chi square and have used this test anywhere in your application**

Chi Square test is another kind of statistical hypothesis testing to test whether relationship exists between categorical value. If P Value is <= 0.5 then we will reject the null hypothesis.

* + - Chi Square statistics compare the tallies or counts of categorical response between to independent groups
    - Formula **X\*\*2 = Sum of((Observed Value – Expected Value)\*\*2/Expected Value)**

**Q12. Can we use Chi square with Numerical dataset? If yes, give example. If no, give Reason?**

Yes, Chi Square test can be used for the Numerical dataset also. (Except Percentage, Proportions, Means, etc.)

* + - Ex :- Count of car having the person.

**Q13. What do you understand by ANOVA Testing?**

ANOVA test is another kind of statistical hypothesis to test the relationship between One numerical feature and more than one categorical feature having with more than 2 categories. If P Value is <=0.5 then we will reject the null hypothesis.

**Q14. Give me a scenario where you can use Z test and T test.**

Z test is statistical hypothesis which is used to determine that where two sample mean calculated are different in case the standard deviation is given, and sample size is large (>30). T test is used where standard deviation is not available and sample size is less (<30).

**Q15. What do you understand by inferential Statistics?**

Inferential statistics is the statistical method to take the random sample data from the population to make conclusions about entire populations.

**Q16. When you are trying to calculate Std Deviation or Variance, why you used N-1 in Denominator? (Hint: Basel Connection)**

Calculation for both sample standard deviation and sample variance contains little bias comparison to the population standard deviation and variance, So when you divide both from n-1 instead n, results would be near to accurate.

**Q17. How you can define Machine Learning?**

Machine learning is a form of AI, that teaches computers to think in a similar way like human do, Learning and improving from the experience, it works by exploring data and identifying patterns and involves minimum human intervention.

**Q18. What do you understand Labelled training dataset?**

Labelled data is used as input for supervised machine learning model, where we divide the dataset into training and testing dataset. Training dataset is used to train the model, So training dataset size should be larger than the test dataset for better performance and accuracy.

**Q19. What are 2 most common supervised ML tasks you have performed so far?**

Logistic Regression and Random Forest.

**Q20. What kind of Machine learning algorithm would you used to walk robot in various unknown area?**

Reinforcement Learning, where machine can take a decision in sequence.

**Q21. What kind of ML algo you can use to segment your user into multiple groups?**

Unsupervised machine learning is the best option to segment the users in multiple groups. Most popular algorithm under this umbrella is – K-Means Clustering, DBSCAN.

**Q22. What type of learning algo realised on similarity measure to make a prediction?**

Multiple learning algorithm like K-nearest neighbor’s , DBSCAN etc use similarity measure to make predictions.

Distance measures are the fundamental principals of classification, which measures the dissimilarity between given data samples.

Algorithms has multiple methods to measure distance Exp. Euclidean, Manhattan, Chebychev, Minkowski.

**Q23. What is out of core learning?**

Set of Algorithms working with data that can not fit into the memory of single computer, but that can easily fit into some data storage such as local hard disk.

**Q24. Can you please give 1 example of hyperparameter tuning wrt some classification algorithm?**

One very popular method to tune hyperparameter is GridSeachCV from the Scikit Learn library. This method tries every possible combination of each set of hyper-parameters, using this method, we can get the best set of values in parameter to achieve high accuracy and performance.

**Q25. What is out of bag evaluation?**

Out of bag is validation process for the random forest model, where small set of records has randomly distributed to different decision tree model for training and testing. So during these process the record set which is not available in training dataset has been validated on testing process by the different decision trees and then majority of vote count as correct result.

**Q26. Let’s Suppose I have trained 5 diff model with same training dataset & all of them have achieved 95% precision. Is there any chance that you can combine all these models to get better result? If yes, How? If no, Why?**

We can use clustering algorithms to cluster the dataset and pass that every cluster to different algorithm to get better accuracy.

**Q27. What do you understand by Gradient decent? How will you explain Gradient decent to a kid?**

Gradient decent is optimization algorithms to identify the slop in all directions at any point to reach the minimum point.

In layman word, If we stand top in hill and wanted to go down bottom of the hill, then we have to repeat some of the steps to reach bottom point.

* + - Start from any point on Hill.
    - Check all the direction to determine where we might be able get descend.
    - Take a step in that direction.
    - Repeat the steps until reach the minimum point. When you reached the point that doesn’t make any difference, would be minimum point.

**Q28. Can you please explain diff between regression & classification?**

* + Regression :- Predict continuous value ex. Salary, age, price. In regression, we try to find best fit line, which can predict output more accurately. Used to solve regression problem like weather prediction, house prediction. It can further divide in to Linear or non-linear Regression.
  + Classification :- Predict/classify discreate value ex. Male/Female, True/False,Spam/Not Spam. Try to find decision boundary, which can divide dataset into different classes. It used to solve classification problem. It can further divide into to binary or multi class classifier.

**Q29. Explain a clustering algorithm of your choice.**

* + We below most popular clustering algorithm.
    - K-Means Clustering
    - Hierarchical Clustering
    - DBSCAN (Density Based Clustering and Application with Noise)

DBSCAN :- It is density based clustering, which has 2 most important hyperparameter, 1) distance epsilon € 2) Minimum Point. Distance epsilon works as radius of circle from any datapoint and if under that circle minimum point available then that would have a core point and within epsilon distance all the point comes in same cluster. Like the similar way process would repeat for all the points. If any point which does not cover the minimum point within his epsilon distance, then that point count as noise / outlier point.

Advantage :- As per the process mechanism, doesn’t require pre-set number of cluster at all. It also identifies noise / outlier data points. It can be find arbitrarily sized and arbitrarily shaped cluster quite well.

Disadvantage :- It doesn’t perform well if low / varying density and if high-dimensional data.

**Q30. How you can explain ML, DL, NLP, Computer vision & reinforcement learning with example in your own terms?**

* + Reinforcement learning is the training of machine learning model to make a sequence of decisions. The agent leans to achieve a goal in an uncertain, potentially complex environment. In Reinforcement learning an artificial intelligence face a game like situation. Exp Autonomous / Self driving Car.