

Detail Project Report

Sentiment analysis

Dinesh Kumar Naik | Ticket Assignment | 03-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:

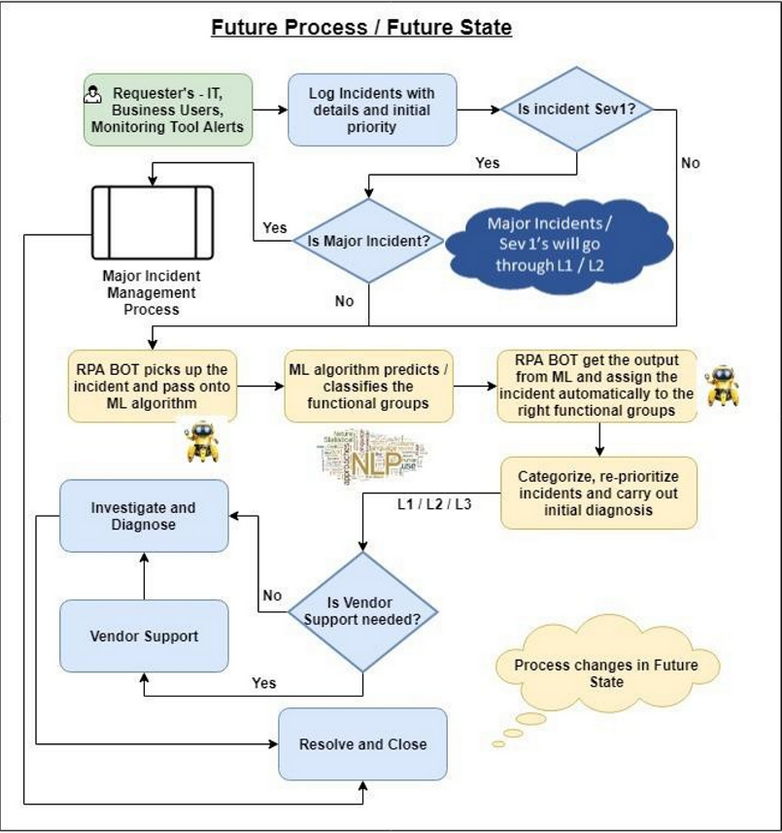
Manual ticket assignment is time consuming and require human efforts, as it requires to read all the ticket description carefully. There may be mistake due to human error and resource consumption is carried out ineffectively because of misaddressing.

## Value of the Project:

* Create machine learning model that can predict / classify the incidents to the right functional group with more than 75% accuracy.
* Integrate the machine learning model with Robotic Process Automation (RPA) BOT to automate the incident assignment.
* L1 Or L2 team manual effort in assigning incident to L3 teams will be eliminated / reduced and they can now focus on higher value-added activities.
* We can demonstrate the success of Machine Learning through this project and be able to take up multiple projects / initiatives.

## System Architecture:

* Functional Flow.



**Above figure shows, how the new incident Management process will look like once the IT Service Management Tool is integrated with RPA and Machine Learning Model.**

**Storage:** Oracle 12g.

**Cloud: IBM Watson Cloud:**

**API:** 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: List of features**

|  |  |
| --- | --- |
| Feature Name | DataType |
| Incident Number | String |
| RaisedBy | String |
| Short Desc | String |
| Description | String |
| Priority | String |
| Created Date | Date |
| Assigned To | String |
| LineOfBusiness | String |
| Resolved | Date |
| Assignment Group | String |

* **Data Pre-Processing:**
  + 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.
  + **Feature Extraction:** The classifier and learning algorithms can not directly process the text documents in their original form, as most of them expect numerical feature vectors.

**Technique :**

* + Count Vector (BOW – Bag of Word)
  + TF-IDF
  + Word2Vec
  + Label Encoding – Target Label

Note :- Count Vector chosen as final feature extraction technique.

* **Algorithms:**
  + Multinomial Naïve Bayes
  + Gradient Boosting Classifier
  + SGD Classifier

After running basic model and simulated various features in order to finalize the feature that will be implemented with other classifiers.

**Hosting:**

**Dashboarding:** OBIEE

**Monitoring:**

**Hypercare:**

**Question Answers:**

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

The project is called classification of ticket allocation.

This is inhouse project to Integrate the machine learning model with Robotic Process Automation (RPA) BOT to automate the incident assignment.

We have ICD tool for the IOC Partner to raise there tickets and L1 team read all the tickets to read and assign the same to the respective functional user.

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

We do receive around 300 tickets 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 classification problem.

|  |  |
| --- | --- |
| Feature Name | DataType |
| Incident Number | String |
| RaisedBy | String |
| Short Desc | String |
| Description | String |
| Priority | String |
| Created Date | Date |
| Assigned To | String |
| LineOfBusiness | String |
| Resolved | Date |
| Assignment Group | String |

**Q4. What technique for data preprocessing, could you please explain the method?**

Count vectorization used for the data preprocessing.

Count Vectorization (Bag of Word) transform given text into vector on the basis of the frequency (count) of each word that occurs in the entire text.

**Q5. Why you use count vectorization?**

Count vectorization doesn’t change the meaning of the sentence and it includes the most of the word to transform in vector. This technique the highlight the word which occurs multiple time. Generally length of description of the issue is short and bow is better, when the no of sentence is less.

**Q6. What is other techniques in NLP to process the sentence, Please explain in detail?**

We have other technique as TF-IDF (Term Frequency – Inverse Document Frequency) and Word2Vec.

* **TF-IDF** – Full form is Term Frequency – Inverse Document Frequency. This process also transforms the sentence to words like the BOW, TF-IDF gives the importance of uncommon words. TF IDF score (float) the words based on the below formula.

TF = No of repetitive words in sentence / No of words in sentence.

IDF = Log(No of sentence / No of sentence containing the words)

* **Word2Vec** – In this model, each word is basically represented as a vector of 32 or more dimension instead of single number. Here the semantic information and relation between different words is also preserved. Ex. Man-Women, King-Queeen.

**Q7. What is stopwords and why we use in NLP data preprocessing.**

Stopword is the word that is commonly used in sentence, that they carry very little useful information. We wanted to focus on important information on sentence, so we remove the stopwords.

**Q8. What is stemming and lemmatization.**

Both methods generate the root word from the inflected words.

|  |  |
| --- | --- |
| **Stemming** | **lemmatization** |
| Stemming is faster | Slower |
| Rule Based | Dictionary Based |
| It produce the Non Existence of word | Produce meaning full word |
| Prefer to use when word meaning is not important. | Prefer to use when word meaning is important |
| Exm - Studies - Studi | Exm - Studies - Study |

**Q9. What is Multinomial Naïve Bayes and why to use in NLP, Please explain.**

Multinomial naïve bayes uses term frequency, the number of times a given term appears in a document. This classifier is suitable for classification with discreate features (e.g. word counts for text classification). Naïve bayes give better performance on text data.

**Q10. What is the difference between Bernoulli, Multinomial and Gaussian Naïve Bayes?**

**Bernoulli Naïve Bayes : -** When all the features are binary, such that they take only two values 0 or 1. Exp. 0 can represent “ word does not occur in the document” and 1 represent the occurrence.

**Multinomial Naïve Bayes : -** When all the features have discrete data, Exp. Move ratings 1 to 5 as each raring has certain frequency. Word count in text.

**Gaussian Naïve Bayes : -** When all the features have continuous data, Exp. Iris dataset features are sepal width, sepal length, petal width, petal length. We can’t represent features in terms of their occurrences. This means data is continuous.

**Q11. What is the Gradient boosting classifier?**

Boosting is method of converting weak learners into strong learners. In boosting each new tree is a fit on a modified version of the original data set.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Exp | Degree | salary | y hat (Avarage of B Colums) | Residual R1 | Residual R2 | R3 |
| 2 | BE | 50K | 75 | -25 | -23 |  |
| 3 | Maths | 70K | 75 | -5 | -3 |  |
| 5 | Masters | 80K | 75 | 5 | 3 |  |
| 6 | Phd | 100K | 75 | 25 | 20 |  |

R2 value come after passing the independent value (Exp. Degree) to the decision tree.

When we pass the independent value to calculate R3 value, it would be overfitting problem, so we would penalize to multiply alpha value with the output from the decision tree.

Alpha value would be between 0 to 1. And formula is.

F(x) = h0(x) +alpha1 \* h1(x) + alpha2 \* h2(x) + ……..+ alphan \* hn(x)

= sum of alphan \* hi(x)

**Q12. What is the SGD classifier?**

Stochastic Gradient Descent (SGD) is simple efficient optimization algorithms used to find the value of parameter/coefficients of functions that minimize a cost function. This is also called as mini batch gradient descent, as we pass the data points in batch. It would have noise on data point as we use k batch size of datapoint.

In Gradient descent, we pass all the datapoint in one go, so it uses huge memory to process the datapoint and it perform slow.

**Q13. What is cost function?**

A cost function is the measure of how wrong the model is in terms of its ability to estimate the relationship between the independent and dependent variable. It helps the learner to correct / change behavior to minimize the mistakes.

**Q14. What was the challenge you face during implementation of this model, and how you mitigate that challenge?**

IOC partner (distributor, stockiest , dealer) raise the tickets in ICD tool. Some of the partner 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.

**Q15. How were you maintaining the failure cases?**

If any ticket assigns to wrong functional group, we do analysis that ticket and if any identifier word missed in model dictionary or required any logic modification, we do correction for next model training.

**Q16. What was the expected success target for your model, and what is the current success %?**

Expected target was around 70 to 80% and current success rate is between 78 to 85%.

**Q17. Have you used any event scheduler?**

Yes, we have used BOT to pass new incident to our process for assignment.

**Q18. What is the frequency to retrain your model?**

It depends based on the performance measurement but within 3-4 weeks, we do re-train the model.

**Q19. What were your roles and responsibilities in the project?**

My responsibility to design the model, preprocessing the data, train the model, analysis the failed cases, Providing the QA Support.

**Q20. In which technology you are most comfortable?**

I have worked in Machine learning and NLP. Exploring on Deep Learning and Computer vision.

**Q21. What do you understand by right skewness, Give example?**

* + Right skewed data has long tail on right side, it is called as positive skew distribution also as long tail is in positive direction line. Mean is also on the right of the median.

Exp : How many people older than 60 lived in there households.

**Q22. What is difference between Normal distribution and Std Normal Distribution and Uniform Distribution?**

* + All three are probability distribution, where shape (bell curve) of the Normal and Std Normal distribution is same and when you divide the shape from the mean point, it looks the mirror image of each other, only difference is Std Normal distribution mean (meu) would have 0 and standard deviation has 1 value but for normal distribution mean value has different than 0. Normal distribution known as Gaussian distribution.
  + Shape of the uniform distribution looks like rectangle, so in the distribution probability of x is constant.

**Q23. What is different kind of Probabilistic distributions you heard of?**

* + Bernoulli Distribution
  + Uniform Distribution
  + Binomial Distribution
  + Normal Distribution
  + Poisson Distribution
  + Exponential Distribution

**Q24. What do you understand by symmetric dataset?**

* + Symmetric dataset is normally distributed in shape, When we visualize through histogram it looks mirror image either side from the middle. For this dataset mean, mode and median would be on same point.

**Q25. In your last project, were you using symmetric data or Asymmetric Data, if its asymmetric, what kind of EDA you have performed?**

* + Asymmetric dataset is not normally distributed, it mean it would have either left/right skewed. We can reduce the skewness applying the below Transformation method.
    - Log Transformation
      * Most used method
      * Easily done by np.log
      * Data should not have null value.
      * Handle values at 0
* Box Cox Transformation
  + Data must be positive
  + Imported from scipy
* Square root Transformation
  + Generally not used as taking square root shortens the range of variable.
  + Can be applied via np.sqrt().

**Q26. Can you please tell me formula for skewness?**

* + Different distribution has different formula to calculate skewness but General formula to calculate the skewness is:
    - Skew = 3\*(Mean-Median)/Standard Deviation.

**Q27. What is difference between Normal distribution and Std Normal Distribution and Uniform Distribution?**

* + All three are probability distribution, where shape (bell curve) of the Normal and Std Normal distribution is same and when you divide the shape from the mean point, it looks the mirror image of each other, only difference is Std Normal distribution mean (meu) would have 0 and standard deviation has 1 value but for normal distribution mean value has different than 0. Normal distribution known as Gaussian distribution.
  + Shape of the uniform distribution looks like rectangle, so in the distribution probability of x is constant.

**Q28. What is different kind of Probabilistic distributions you heard of?**

* + Bernoulli Distribution
  + Uniform Distribution
  + Binomial Distribution
  + Normal Distribution
  + Poisson Distribution
  + Exponential Distribution

**Q29. What do you understand by symmetric dataset?**

* + Symmetric dataset is normally distributed in shape, When we visualize through histogram it looks mirror image either side from the middle. For this dataset mean, mode and median would be on same point.

**Q30. What do you understand by statistical analysis of data, Give me scenario where you have used statistical analysis in last projects?**

* + Statistical analysis is the science of collecting data and analysis the pattern and trend. Below is some of the analysis process.
    - Summarise the data based on the data visualization.
    - Measures of the location to finding mean.
    - Measure of the spread to finding standard deviation.
    - Make future predictions based on the past behaviour.
    - Hypothesis testing.