"Sentiment Analysis of Twitter data Using NLP"

A Project Report

Submitted in fulfillment for the award the degree of

ADVANCED PROFESSIONAL CERTIFICATE PROGRAMME

IN DATASCIENCE AND MACHINE LEARNING

Submitted to

E&ICT, IIT GUWAHATI & JARO EDUCATION

Submitted By:		
Mr. Sachin S. Malode *M.E.(Digital Electronics)*	*Mr. Ganesh Malgunde* *M.Tech (Space Science)*	*Mr. Srikanth Bukya* *B.Tech (Mechanical)*

Group Name: GROUP 15

Project Guide:

Prof. Sarveshwaran Rajagopal

(Data scientist & Trainer)

Team & Roles:

Team Overview:

Our team comprises individuals with complementary skills in data science, software engineering, and domain expertise, united by a common goal of leveraging AI/ML technologies to solve complex challenges. With a collaborative mindset and a shared passion for innovation, we are committed to delivering high-quality solutions that address the needs of our stakeholders effectively.

Team Strengths:

- **Interdisciplinary Expertise**: Our team combines expertise from multiple disciplines (electronics, space science, mechanical), enabling us to tackle diverse aspects of the project, from data analysis to software development.
- **Effective Communication**: We prioritize open communication and regular collaboration to ensure alignment with project goals and foster creativity and innovation.
- **Adaptability and Resilience**: With a flexible approach and the ability to adapt to evolving requirements, we navigate challenges effectively and iterate on our solutions to achieve optimal outcomes.
- **Commitment to Quality & Deadline**: We are dedicated to delivering solutions of the highest quality, adhering to best practices in AI/ML, software engineering, and project management.

Project Approach:

Our approach involves iterative development cycles, starting with thorough problem analysis and requirements gathering. We emphasize prototyping and experimentation to explore different AI/ML techniques and refine our solutions based on feedback. Regular milestones and checkpoints ensure steady progress and allow for course correction as needed.

Team Members:

	Member- 1	**Member- 2**	**Member-3**
Name:	Mr. Sachin S. Malode	Mr. Ganesh Malgunde	Mr. Srikanth Bukya
Role:	Data Scientist and Project Lead	Software Engineer and Developer	Domain Expert and Researcher

- **[**Mr. Sachin S. Malode**]** - Data Scientist and Project Lead

- *Role*: As the project lead and data scientist, I bring expertise in machine learning algorithms, data pre-processing, and model evaluation. My responsibilities include problem formulation, platform selection, algorithm selection, and ensuring the technical integrity of our solutions. I have experience with various machine learning frameworks and a strong background in statistical analysis. Also scheduling meetings, taking feedback from team, guiding time to time, Research & helping in report writing. testing and evaluation of final outcome, these responsibilities carried out successfully during project building & implementation.

- **[**Mr. Ganesh Malgunde**]** - Software Engineer and Developer

- *Role*: As the software engineer and developer, is responsible for implementing scalable and efficient AI models into production-ready software solutions. With expertise in software development and system architecture, they ensure seamless integration of machine learning algorithms with existing infrastructure. Their skills in coding, debugging, and optimization are invaluable for delivering robust and reliable software solutions. Writing code, implementing different algorithm, training the model, connect with team on regular basis to clear difficulties in implementation , selecting model, algorithms ,deciding flow of implementation etc this responsibilities carried out successfully during project building & implementation.

- **[**Mr. Shrikanth Bhukya**]** - Domain Expert and Researcher

- *Role*: Bringing domain-specific knowledge and research expertise, plays a crucial role in understanding the context and requirements of the project. With a deep understanding of the industry or problem domain, they provide valuable insights for refining project objectives and designing tailored AI/ML solutions. Their research skills aid in staying updated with the latest advancements in AI/ML techniques relevant to our project. Read research papers & articles, helping team member in selecting libraries & algorithms in accordance with project requirement and give suggestion to team for implementation. Preparing project report, documentation, code collection is final step of project is completed with the help of team.

Problem Statement:

This is the sentiment140 dataset. It contains 10,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment.

Content

It contains the following 6 fields:

1. **target**: the polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)
2. **ids**: The id of the tweet
3. **date**: the date of the tweet
4. **flag**: The query (*lyx*). If there is no query, then this value is NO_QUERY.
5. **user**: the user that tweeted (eg., *robotickilldozr*)
6. **text**: the text of the tweet (eg., *Lyx is cool*)

Task:

Task1: Clean the text – Remove special characters

Task2: Use embedding technique to convert the word to vectors

Task 3: Tokenize the text

Task 4: Built model

Task 5: Evaluate the model and print the confusion matrix

Objective:

To find the sentiment of 10,000 tweets, extracted using twitter API, analyzing the words from those tweets by building a model.

A Twitter sentiment analysis determines negative, positive, or neutral emotions within the text of a tweet using NLP and ML models. Sentiment analysis or opinion mining refers to identifying as well as classifying the sentiments that are expressed in the text source. Tweets are often useful in generating a vast amount of sentiment data upon analysis. These data are useful in understanding the opinion of people on social media for a variety of topics

Why is Twitter Sentiment Analysis Important?

1. **Understanding Customer Feedback:** By analysing the sentiment of customer feedback, companies can identify areas where they need to improve their products or services.
2. **Reputation Management**: Sentiment analysis can help companies monitor their brand reputation online and quickly respond to negative comments or reviews.
3. **Political Analysis**: Sentiment analysis can help political campaigns understand public opinion and tailor their messaging accordingly.
4. **Crisis Management:** In the event of a crisis, sentiment analysis can help organizations monitor social media and news outlets for negative sentiment and respond appropriately.
5. **Marketing Research:** Sentiment analysis can help marketers understand consumer behaviour and preferences, and develop targeted advertising campaigns.

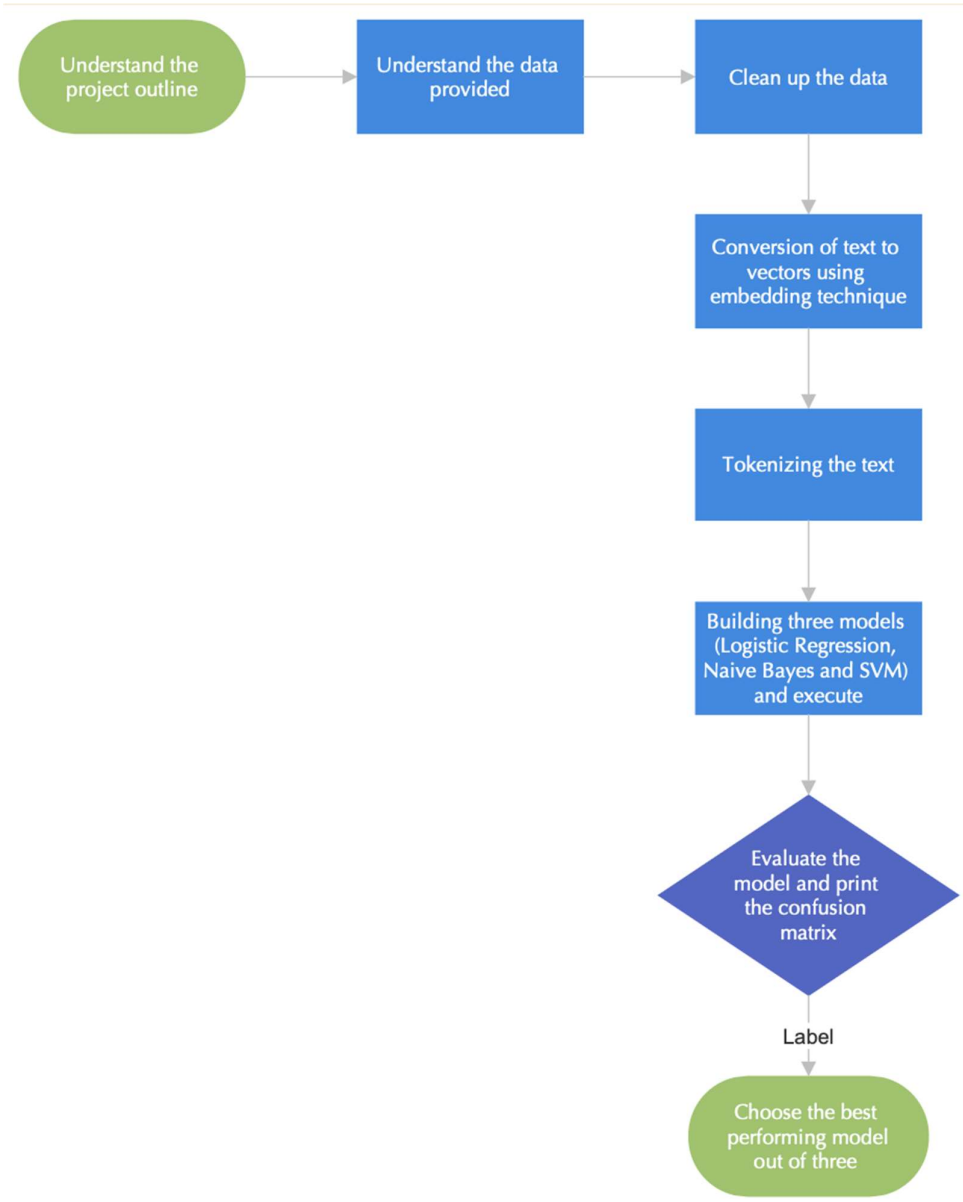
How to Do Twitter Sentiment Analysis Dataset?

We aim to analyse twitter sentiment analysis dataset using machine learning algorithms, the sentiment of tweets provided from the **Sentiment140 dataset** by developing a machine learning pipeline involving the use of three classifiers (**Logistic Regression, Bernoulli Naive Bayes, and SVM**) along with using. The performance of these classifiers is then evaluated using **accuracy** and **F1 Scores**.

IMPLEMENTATION STEPS:

1. **Problem Definition and Scope**:
    - Collaborated with the team to define the problem statement and scope of the sentiment analysis project.
    - Conducted thorough research to understand the domain and potential applications of sentiment analysis.
2. **Data Acquisition and Preprocessing**:
    - Identified relevant datasets suitable for sentiment analysis, considering factors such as data size, diversity, and quality.
    - Developed data preprocessing pipelines to clean and prepare the raw text data for analysis.
    - Applied techniques such as tokenization, stop-word removal, and stemming/lemmatization to standardize the text data.
3. **Feature Engineering**:
    - Extracted meaningful features from the text data to represent sentiment-related information effectively.
    - Utilized techniques such as word embeddings (e.g., tfidf, Word2Vec) to represent words and phrases in a numerical format.
    - Conducted exploratory data analysis (EDA) to gain insights into the distribution of sentiment labels and identify potential patterns in the data.
4. **Algorithm Selection and Implementation**:
    - Researched various machine learning and deep learning algorithms suitable for sentiment analysis tasks, such as Naive Bayes, Support Vector Machines (SVM), and Recurrent Neural Networks (RNNs)
    - Implemented and fine-tuned the selected algorithms to achieve optimal performance in terms of accuracy and computational efficiency.
    - Utilized libraries such as scikit-learn, for model development and training.
5. **Model Training and Evaluation**:
    - Designed appropriate training-validation splits.
    - Conducted hyperparameter tuning to optimize model performance and generalization ability.
    - Evaluated models using relevant metrics such as accuracy, precision, recall, F1-score.
6. **Testing and Validation**:
    - Conducted rigorous testing of the trained models on unseen test data to assess their real-world performance.
7. **Documentation and Reporting**:
    - Documented the entire data science pipeline, including data preprocessing steps, feature engineering techniques, model architectures, and experimental results.
    - Prepared comprehensive reports and presentations summarizing the methodology, findings, and insights gained from the sentiment analysis project.
8. **Collaboration and Communication**:
    - Actively collaborated with other team members, to ensure alignment with project objectives and timelines.
    - Engaged in regular meetings and discussions to share progress updates, discuss challenges, and brainstorm potential solutions.
    - Provided guidance and support to teammates on data-related issues and best practices in machine learning and data science.

Workflow Diagram:



ML Algorithm used in Sentiment Analysis:

Logistic regression:

Logistic regression is a method of modeling the data with discrete outcomes most popularly used for binary outcomes such as true/false, yes/no and so on. Some popular examples of its use include predicting if an email is spam or not spam or if a tumor is malignant or not malignant. It is one of the more common classifiers for binary classification.

It is also used where there are more than two possible outcomes, which is called multinomial logistic regression. However the order of these outcomes have no specific order. This is easier to build and implement machine learning methods. It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.

Naive Bayes algorithm:

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem of probability. This model predicts based on the probability of an object. It assumes the occurrence of one feature is independent of the occurrence of another feature. That's why it is called 'Naive'. The formula for Bayes theorem is given as:

Some popular examples of Naive Bayes algorithm are spam filtration, sentimental analysis. Some popular examples of Naive Bayes algorithm are spam filtration, Sentiment analysis, and classifying articles.

Naive Bayes formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

- P(A|B) – the probability of event A occurring, given event B has occurred
- P(B|A) – the probability of event B occurring, given event A has occurred
- P(A) – the probability of event A
- P(B) – the probability of event B

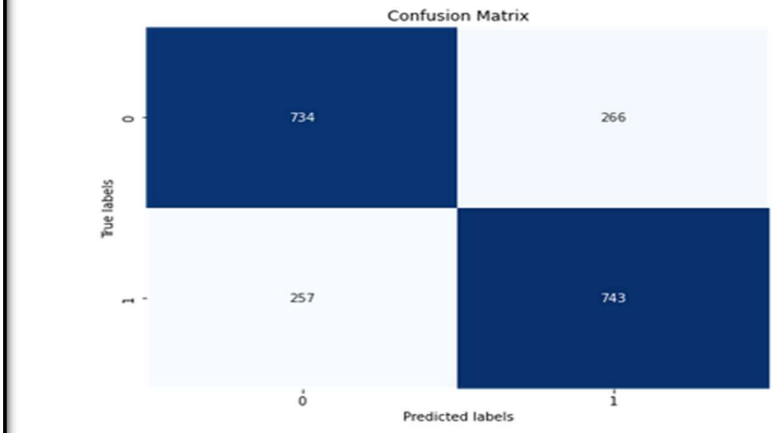
Support Vector Machine (SVM) algorithm:

Support Vector Machine (SVM) algorithm is also a supervised learning algorithm. This is used to classify the data points into different classes in n-dimensional space so that the new data point can be placed in the correct class in the future. SVMs are useful for analyzing complex data that can't be separated by a simple straight line. This is done by using a mathematical trick that transforms data into higher dimensional space, where it is easier to find a boundary. This algorithm uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

OUTPUT:

1. Model Evaluate (Logistic model)

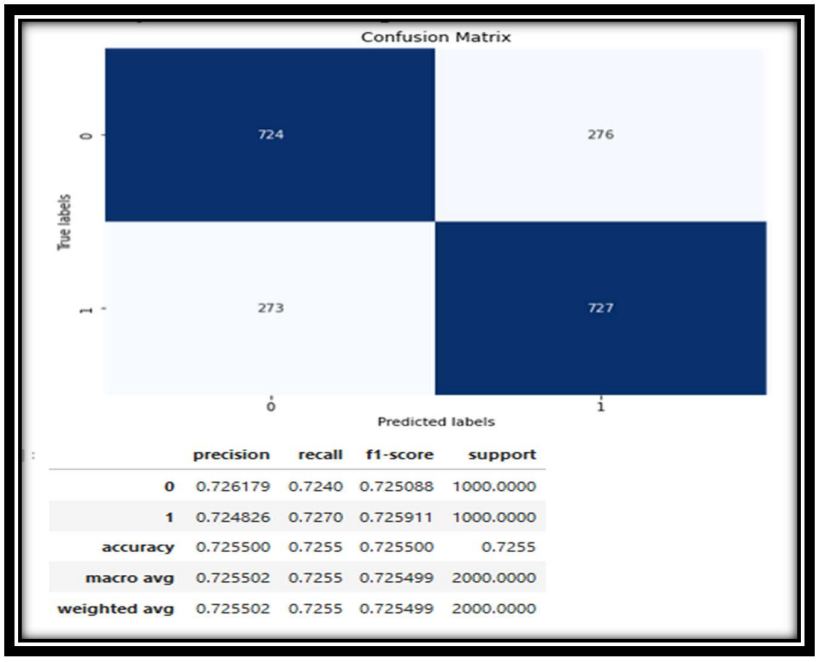
```
In [45]: model_Evaluate(Logistic_model)
```



Out[45]:

	precision	recall	f1-score	support
0	0.740666	0.7340	0.737318	1000.0000
1	0.736373	0.7430	0.739671	1000.0000
accuracy	0.738500	0.7385	0.738500	0.7385
macro avg	0.738519	0.7385	0.738495	2000.0000
weighted avg	0.738519	0.7385	0.738495	2000.0000

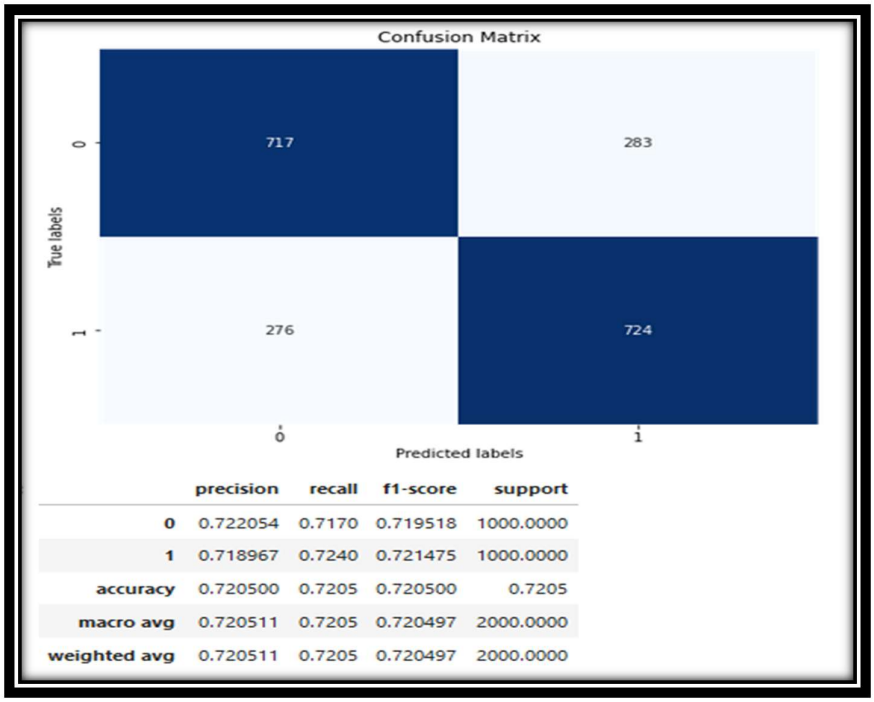
2. Model 2 = Bernoulli Naive Bayes Classifier



:

	precision	recall	f1-score	support
0	0.726179	0.7240	0.725088	1000.0000
1	0.724826	0.7270	0.725911	1000.0000
accuracy	0.725500	0.7255	0.725500	0.7255
macro avg	0.725502	0.7255	0.725499	2000.0000
weighted avg	0.725502	0.7255	0.725499	2000.0000

3.Model 3 = SVM (Support Vector Machine)



	precision	recall	f1-score	support
0	0.722054	0.7170	0.719518	1000.0000
1	0.718967	0.7240	0.721475	1000.0000
accuracy	0.720500	0.7205	0.720500	0.7205
macro avg	0.720511	0.7205	0.720497	2000.0000
weighted avg	0.720511	0.7205	0.720497	2000.0000

Interpretation of results:

Upon evaluating all the models, we can conclude the following details i.e.

Accuracy: As far as the accuracy of the model is concerned, Logistic Regression performs better than Bernoulli Naive Bayes., which in turn performs better than SVM

F1-score: The F1 Scores for class 0 and class 1 are :
(a) For class 0: SVM (accuracy =0.71) < Bernoulli Naive Bayes(accuracy = 0.72) < Logistic Regression (accuracy = 0.73)
(b) For class 1: SVM (accuracy = 0.72) <Bernoulli Naive Bayes (accuracy = 0.72) < Logistic Regression (accuracy = 0.73)

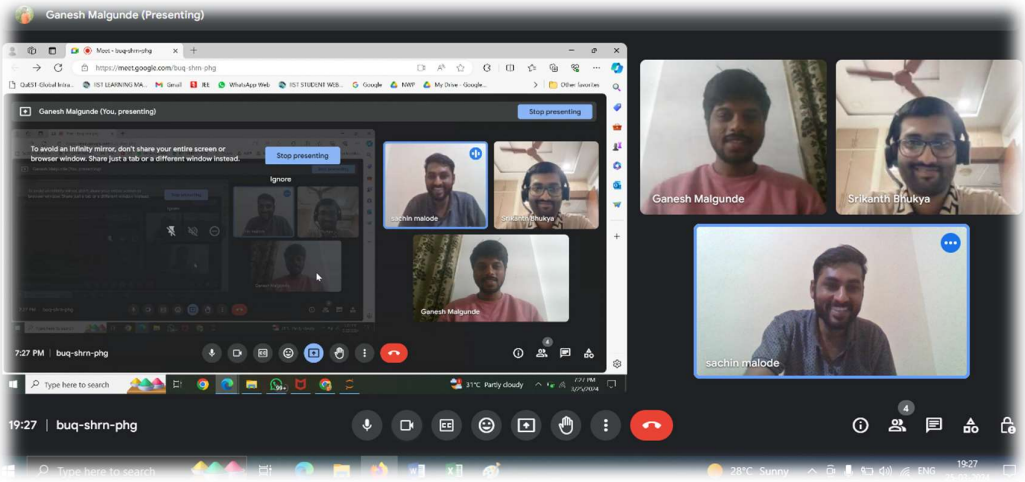
Conclusion:

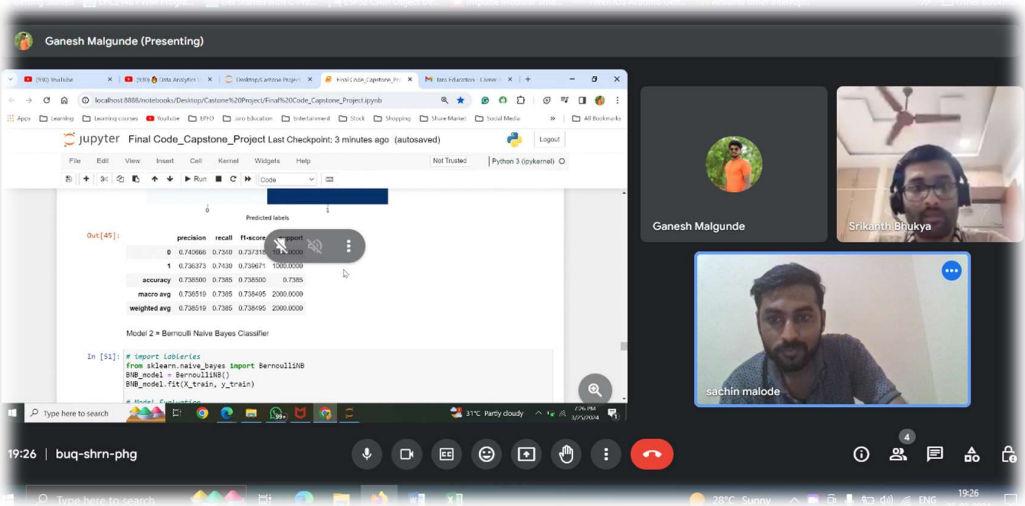
Sentiment Analysis is used to understand public emotions behind people's tweets. Twitter Sentimental Analysis dataset helps us preprocess the data (tweets) using different methods and feed it into ML models to give the best accuracy. On the basis accuracy & F1-score ,We conclude that the Logistic Regression is the best model for the above-given dataset.

References:

1. Sentiment Analysis with NLP on Twitter Data, International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 11-12 July, 2019
2. Social Media Sentiment Analysis On Twitter Datasets, 6th International Conference on Advanced Computing & Communication Systems (ICACCS)
3. Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques, Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)

❖ Glimpses of Discussions & Meetings :





Thank You!