Question2:

Answer2:

To implement this real time decision support system, we would need to collect few metrics which would be helpful to build reliable machine learning model. some of the useful features are

* **User History**: The history of the user who posted the comment can also be a factor in determining the legitimacy of a comment. For example, users who have a history of posting fraudulent comments might be considered less trustworthy.
* **User demographics**: Demographic information such as age, gender, and location can be used to determine the legitimacy of a comment. For example, comments from users in countries where fake comments are prevalent might be considered less trustworthy.
* **comment length**: The number of words in the comment may indicate whether it's real or illigit.
* **Sentiment** : The sentiment expressed in a comment can also be used to determine its legitimacy. For example, a comment that is overwhelmingly positive or negative may be considered suspicious and less trustworthy.
* **Helpfulness**: Comments that include images or videos of the product being commented can be more trustworthy, as they provide tangible evidence of the user's experience. Also comment which includes helpful information about the contents of the book is considered to be legitimate
* **Comments per user**: number of comments per user is useful metric as very high comments can be bot generated and very few comments can be a chance that account is created just to give illegitimate comment.

Python libraries like **nltk** is used for processing comments like removing stop words, tokenising the comments which could be used as input to machine learning model. **Numpy** and **pandas** for data transformation. scikit learn library can be used for building ML models and error metric calculations.

Tensorflow library can be used to build deep learning/ neural networks models.

Supervised machine learning model could be implemented with traning set having tagged legit/illigit comment data.

Machine learning models that can be used for filtering the illigit comments are decision trees, random forests, support vector machines, catboost, Xgboost, Adaboost, Gradient Descent or neural networks

which ML model to choose depends on roc\_auc value of the models when tested. model which gives higher roc\_auc value is best model. Ideally, I would train all the above mentioned model on training data and test it to find best model which gives less False negatives(falsely saying it is legit comment).

Also, hypertuning the parameters is very important, some models give better results when tuned.

My preference of architecture from opensource tools:

Python with Apache spark, SQL database connector, spark SQL could be used to implement opensource decision support system.( would take higher time to implement and not real time scalable compared to cloud tools but budget friendly). Most cloud providers would give no code or less code services which takes lesser time to implement.

Cloud based tools which could be used :

**Azure** : (my preference in cloud tools for decision support system because Azure cognitive services is more matured and gives reliable results when compared to other providers)

**Azure database with My SQL** ->**Azure Cognitive Services** provides pre-built NLP models that can be used to perform sentiment analysis and extract important features from the text of the comments, such as the use of positive or negative words -> **Azure machine learning** service(to build , train and deploy Ml models)

**GCP**:

Cloud SQL-> cloud natural language -> AutoML/vertex AI

AWS:

Aws S3-> Amazon comprehend -> Amazon Sage maker

Some details of machine learning models:

Naive Bayes: This is a simple, probabilistic algorithm that can be used to classify text data based on the frequency of words or phrases in the text. Naive Bayes has been shown to be effective for sentiment analysis tasks and is often used for filtering out spam or illegitimate reviews.

Decision Trees: This is a type of machine learning algorithm that uses a tree-like structure to make predictions based on the values of certain features. Decision trees can be trained on text data by converting the text into numerical features, such as word counts or sentiment scores, and then using these features to make predictions.

Random Forests: This is an ensemble method that combines multiple decision trees to make more accurate predictions. Random forests can be trained on text data and used to classify reviews as legitimate or illegitimate.

Support Vector Machines (SVMs): This is a type of machine learning algorithm that uses linear boundaries to separate data into different classes. SVMs have been shown to be effective for text classification tasks and can be trained on reviews to identify legitimate or illegitimate reviews.

Neural Networks: This is a type of machine learning algorithm that is inspired by the structure and function of the human brain. Neural networks can be used to classify text data by learning to identify patterns and relationships between the words in the text..