**Project 5: Sentiment analysis for marketing – Twitter airline sentiment**

* PURUSHOTHAMAN R, 710021106029, Dept. of ECE, Anna University RC Coimbatore

***Project Title:****Sentiment Analysis for marketing*

**DATASET : Twitter airline sentiment**

**Problem statement explanation:**

* The project's objective is to perform sentiment analysis on customer feedback regarding competitor products.
* By utilizing various NLP methods, the aim is to extract valuable insights into customer sentiments, allowing companies to identify strengths and weaknesses in competing products.
* These insights enable data-driven decision-making, influencing product development, marketing strategies, and ultimately, improving overall customer satisfaction and competitive advantage.
* Continuous monitoring of evolving sentiments is essential for staying responsive to changing customer perceptions.

**DATA WRANGLING :** Data wrangling, also known as data preprocessing or data cleaning, is a critical step in the data preparation process for artificial intelligence (AI) and machine learning tasks. It involves cleaning, transforming, and structuring raw data into a suitable format for analysis and model training

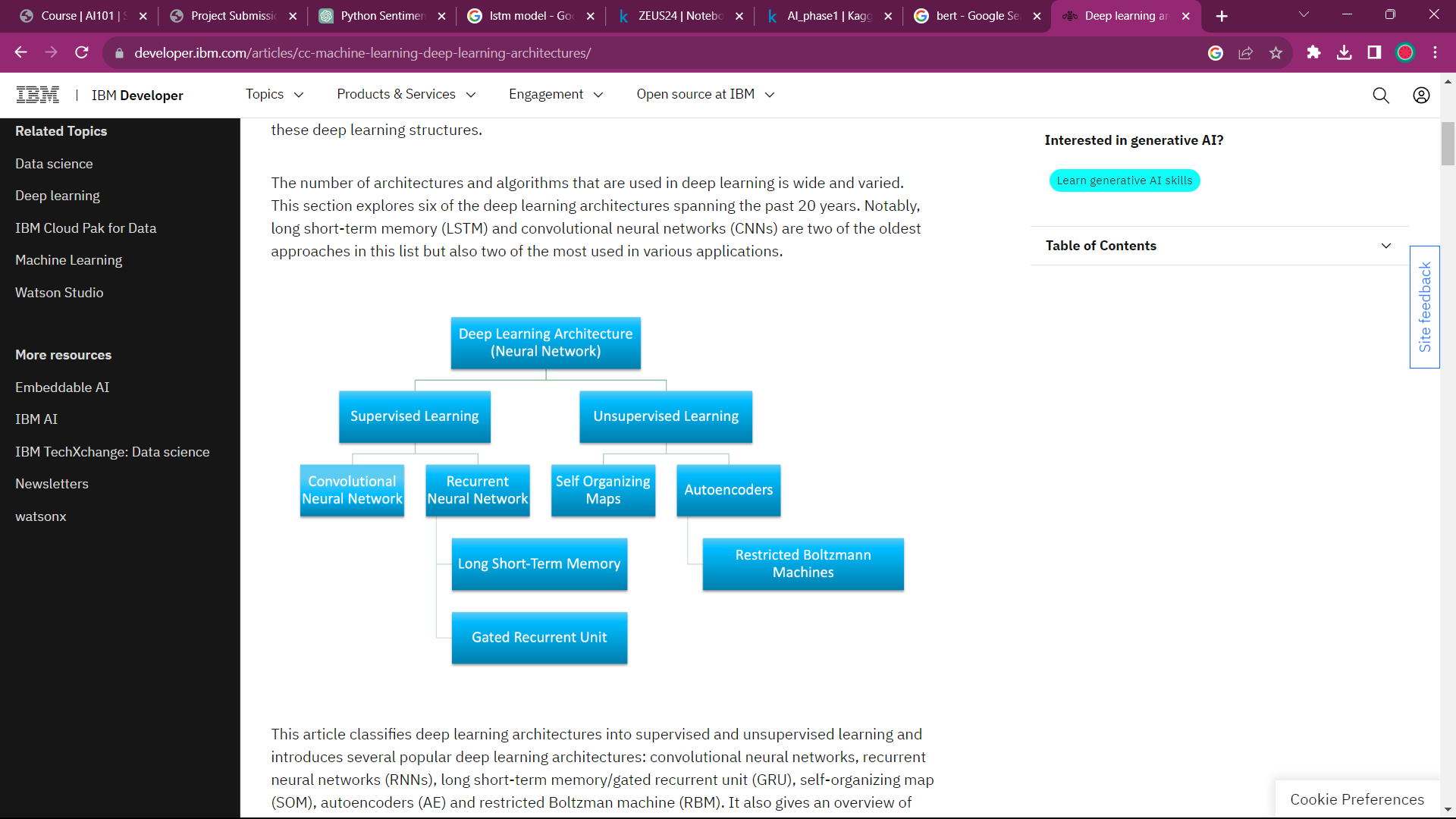
* Data Collection and Retrieval
* Data Cleaning
* Data Transformation
* Data Integration
* Data Reduction
* Data Splitting
* Data Normalization and Standardization
* Data Imbalance Handling
* Data Enrichment
* Data Quality Checks
* Data Visualization
* Data Labeling
* Time Series Data Handling
* Data Storage and Management

***WHAT IS DEEP LEARNING?***

* Deep learning refers to the design and structure of neural networks used in machine learning.
* Deep learning architectures are characterized by having multiple layers, often referred to as deep neural networks.
* These architectures are used for a wide range of tasks, including image and speech recognition, natural language processing, and more.

**UNSUPERVISED LEARNING:** Unsupervised learning refers to the problem space wherein there is no target label within the data that is used for training.  
  
**SUPERVISED LEARNING:** Supervised learning refers to the problem space wherein the target to be predicted is clearly labelled within the data that is used for training.

Deep Learning requires the dataset to be categorized so that we can create new machine learning models that can adapt according to the customer reviews and ratings.



**Random forest algorithm**

* The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees.

* Feature randomness, also known as feature bagging or “the random subspace method” generates a random subset of features, which ensures low correlation among decision trees.

* This is a key difference between decision trees and random forests. While decision trees consider all the possible feature splits, random forests only select a subset of those features.
* By accounting for all the potential variability in the data, we can reduce the risk of overfitting, bias, and overall variance, resulting in more precise predictions.

**Long Short Term Memory**

Long Short Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is specifically designed to handle sequential data, such as time series, speech, and text. LSTM networks are capable of learning long-term dependencies in sequential data, which makes them well suited for tasks such as language translation, speech recognition, and time series forecasting.

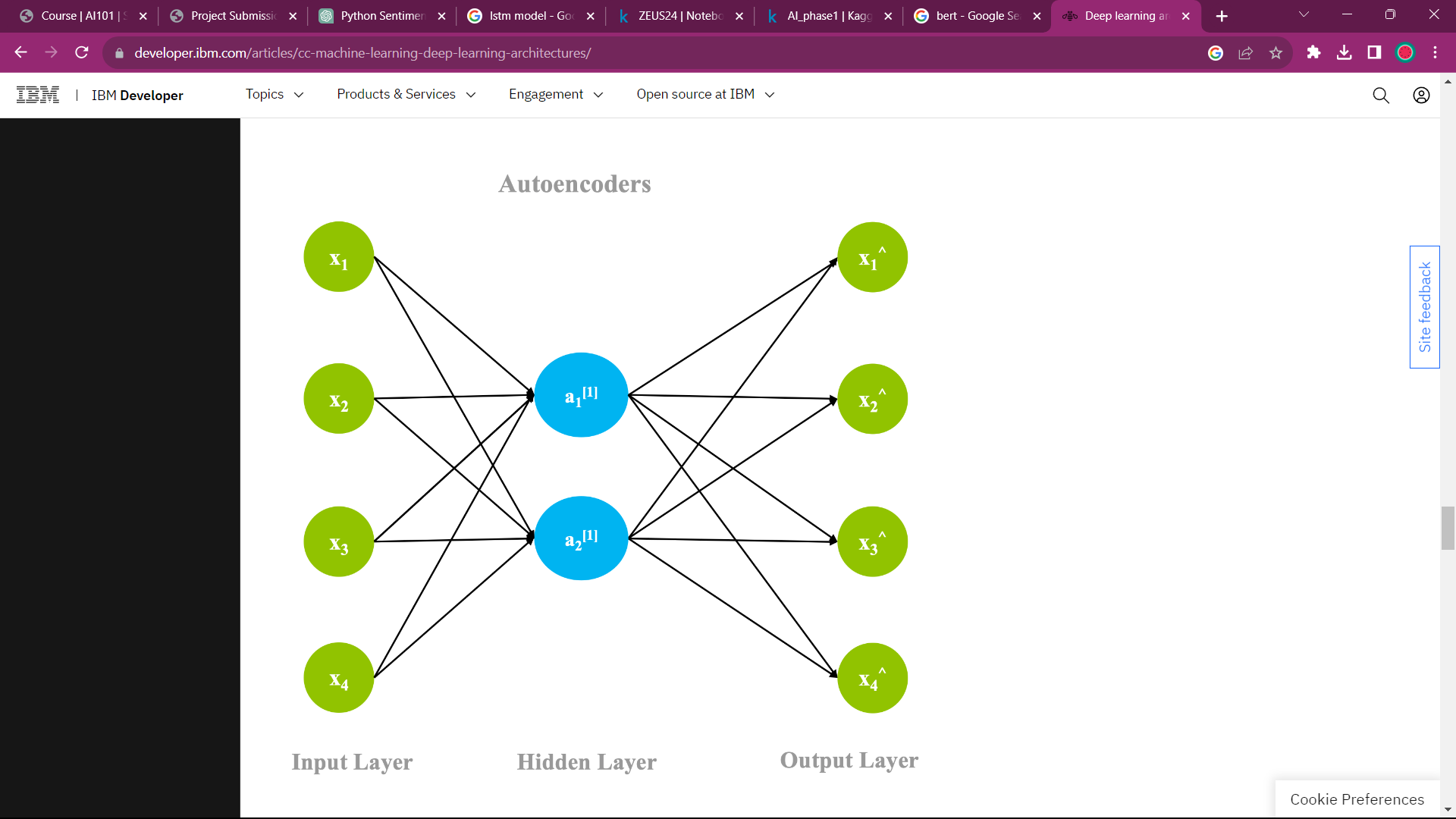
**Advantages of LSTM**

Long-term dependencies can be captured by LSTM networks. They have a memory cell that is capable of long-term information storage.

In traditional RNNs, there is a problem of vanishing and exploding gradients when models are trained over long sequences. By using a gating mechanism that selectively recalls or forgets information, LSTM networks deal with this problem.

**AUTOENCODERS**

* This variant of an *ANN* is composed of 3 layers: input, hidden, and output layers.
* First, the input layer is encoded into the hidden layer using an appropriate encoding function. The number of nodes in the hidden layer is much less than the number of nodes in the input layer. This hidden layer contains the compressed representation of the original input. The output layer aims to reconstruct the input layer by using a decoder function.



**Restricted Boltzmann Machines**

* An RBM is a 2-layered neural network. The layers are input and hidden layers. In RBMs every node in a hidden layer is connected to every node in a visible layer. In a traditional Boltzmann Machine, nodes within the input and hidden layer are also connected. Due to computational complexity, nodes within a layer are not connected in a *Restricted* Boltzmann Machine.

**Bidirectional Encoder Representations**

**1. Pretraining and Fine-Tuning:**

* BERT uses a two-step process: pretraining and fine-tuning.
* In the pretraining phase, BERT is trained on a massive corpus of text data (e.g., Wikipedia) to learn a general language understanding.

**2. Bidirectional Context:**

* BERT is bidirectional, which means it can consider the context of a word by looking at both the left and right context in a sentence.
* This is in contrast to traditional models like LSTMs and traditional transformers, which are unidirectional and only look at the left or right context.

**3. Attention Mechanism:**

* BERT uses self-attention mechanisms, which allow it to weigh the importance of different words in the input text when generating representations.
* Attention heads capture relationships and dependencies between words in the input text.

**RoBERTa**

**Architecture and Pretraining:**

* RoBERTa is based on the transformer architecture, similar to BERT.
* Like BERT, RoBERTa undergoes a pretraining phase on a large corpus of text data, during which it learns to predict masked words in sentences.

**Key Optimizations:**

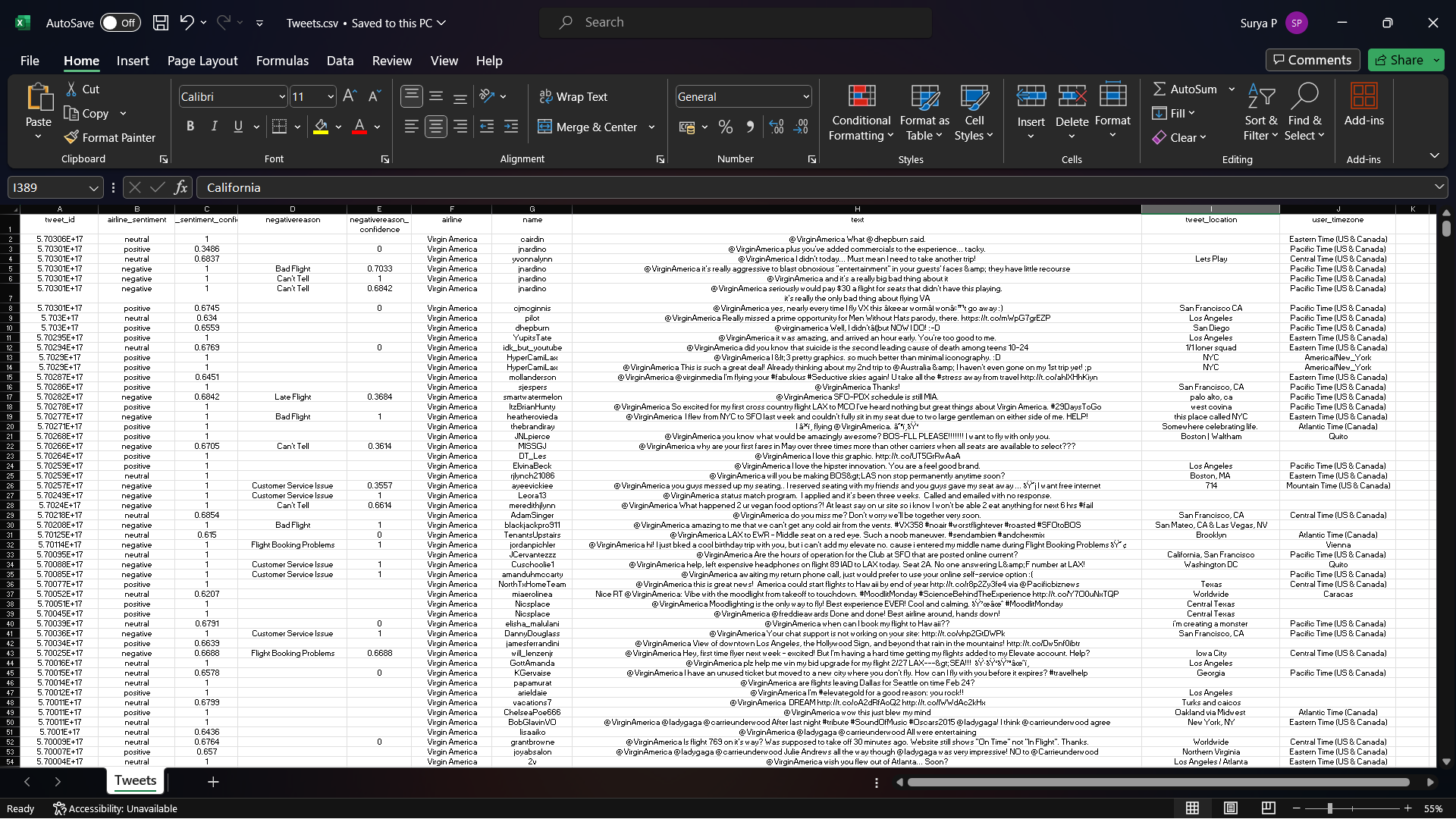
* RoBERTa introduces several key optimizations compared to BERT, including:
  + **Dynamic Masking:** RoBERTa dynamically selects and masks words during pretraining, unlike BERT, which uses static masking. This dynamic masking helps RoBERTa learn more effectively.

**Bidirectional Context Understanding:**

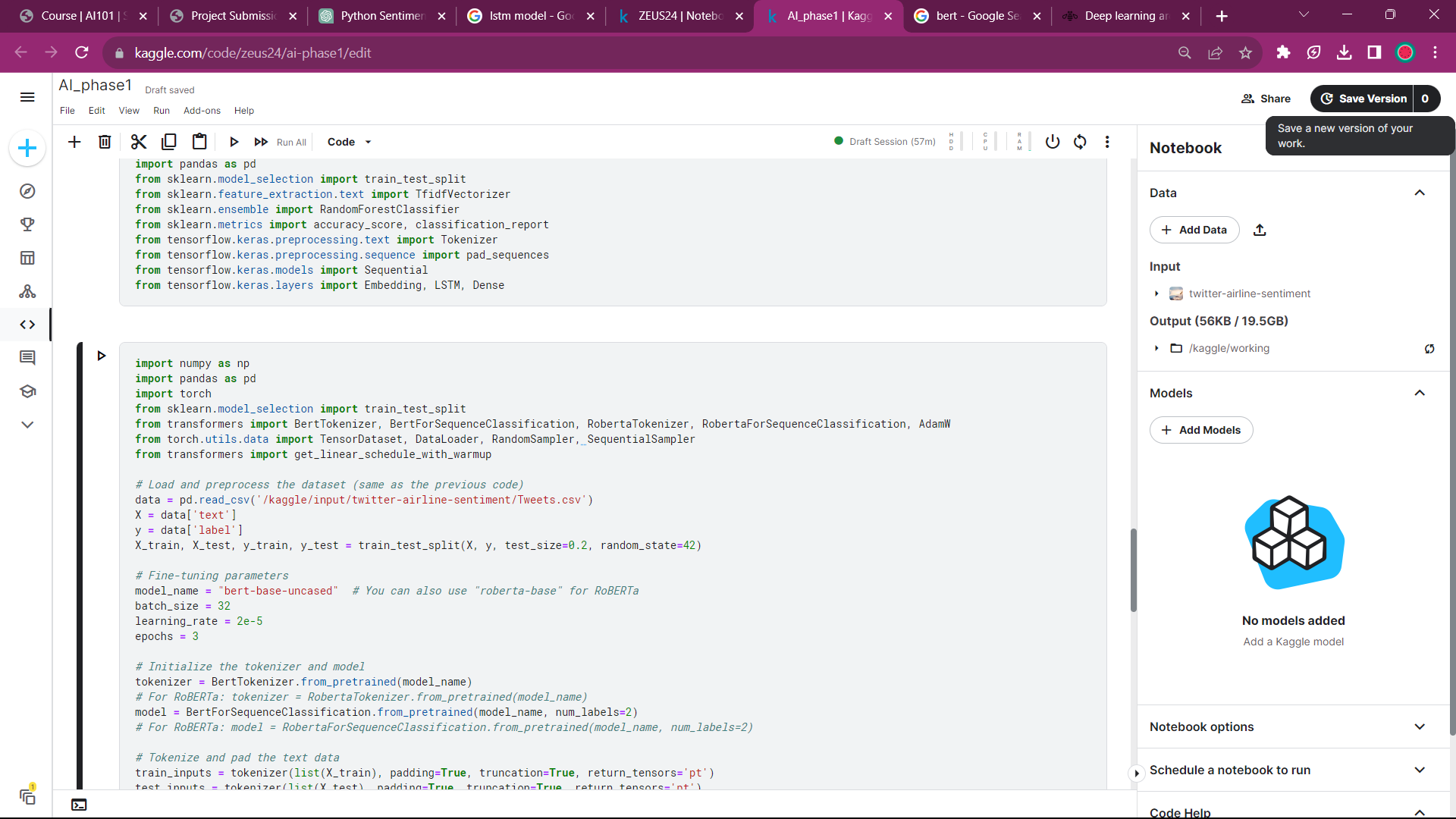
* + RoBERTa, like BERT, is bidirectional in its context understanding. It considers both the left and right context of each word in a sentence, allowing it to understand the full context and semantics.

**Fine-Tuning:**

* + RoBERTa's pretrained model can be fine-tuned on specific NLP tasks. The fine-tuning process is similar to that of BERT and involves training the model on task-specific data.



The dataset is under process and the Machine learning model is being created using the techniques: RoBerta and neural Networks Algorithms



CONCLUSION:

In conclusion, our exploration of serverless architecture for IoT data processing signifies a groundbreaking innovation in the realm of digital transformation. By embracing the serverless approach, we empower organizations to efficiently manage the colossal influx of IoT data, ensuring cost-effective scalability, real-time responsiveness, and unparalleled flexibility. This innovation not only redefines the way we process and utilize data but also opens new frontiers for IoT applications, paving the way for a more connected and data-driven future where businesses and individuals can harness the full potential of the Internet of Things.

Long short-term memory (LSTM) network is a recurrent neural network (RNN), aimed to deal with the vanishing gradient problem present in traditional RNNs. Its relative insensitivity to gap length is its advantage over other RNNs, hidden Markov models and other sequence learning methods.