Chapter (4) With answer

1- What is the difference between binary, multi-class, and multi-label classification?

- If the number of classes is two, it's called binary classification.
- If the number of classes is more than two, it's referred to as multiclass classification.
- In multilabel classification, a document can have one or more labels/classes attached to it

2- Give some applications of text classification. ?

- Content classification and organization
- Customer support
- E-commerce
- language identification
- segregate fake news from real news

3- Describe the pipeline for building text classification systems .?

1. Collect or create a labeled dataset suitable for the task.
2. Split the dataset into two (training and test) or three parts: training, validation (i.e., development), and test sets, then decide on evaluation metric(s).
3. Transform raw text into feature vectors.
4. Train a classifier using the feature vectors and the corresponding labels from the training set.
5. Using the evaluation metric(s) from Step 2, benchmark the model performance on the test set.
6. Deploy the model to serve the real-world use case and monitor its performance

4- Classification can be done without the text classification pipeline, explain how ?

- A simple solution could be to create lists of positive and negative words in English—i.e., words that have a positive or negative sentiment

- Further enhancements to this approach may involve creating more sophisticated dictionaries with degrees of positive, negative, and neutral sentiment of words or formulating specific heuristics (e.g., usage of certain smileys indicate positive sentiment) and using them to make predictions. This approach is called lexicon-based sentiment analysis.

5- Describe with an example the confusion matrix of a classifier. ?

☑ **A confusion matrix** is a table that is used to evaluate the performance of a classifier by comparing the actual and predicted class labels.

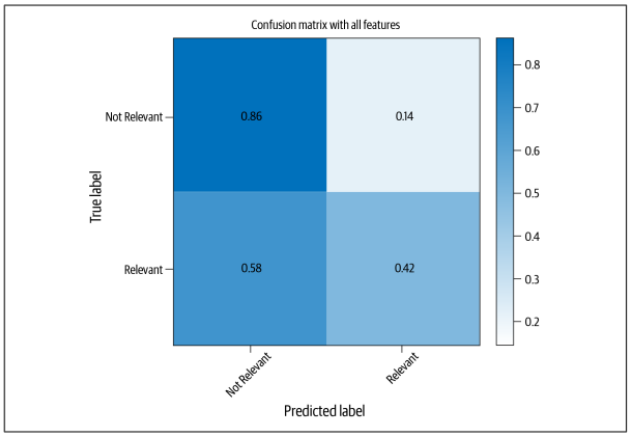

Figure 4-4. Confusion matrix for Naive Bayes classifier

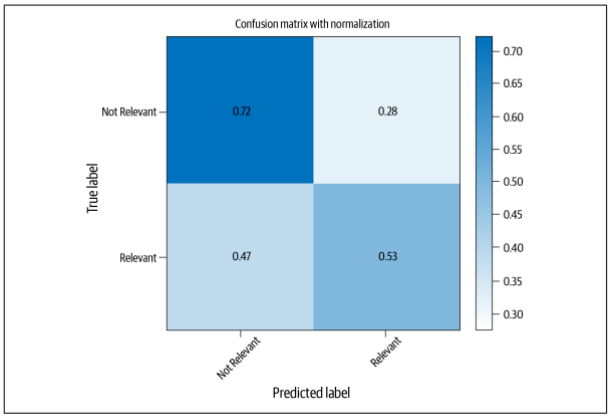

Figure 4-7. Confusion matrix for classification with SVM

6- List the potential reasons for poor classifier performance. ?

☑ Perhaps we need a better learning algorithm

☑ Perhaps we need a better pre-processing and feature extraction mechanism

☑ Perhaps we should look to tuning the classifier's parameters and hyperparameters

7- How to solve class imbalance problem of a dataset ?

☑ Use the right evaluation metrics

☑ Resample the training set

☑ Resample with different ratios

☑ Design your own models

8- What is the difference between generative and discriminative classifiers ?

☑ **generative classifier** → learns the probability of a text for each class and chooses the one with maximum probability.

☑ **discriminative classifier** →that aims to learn the probability distribution over all classes

9- How to use word embeddings as features for text classification ?

☑ Words and n-grams have been used primarily as features in text classification for a long time. Different ways of vectorizing words have been proposed, and we used one such representation in the last section, CountVectorizer.

☑ neural network–based architectures have become popular for “learning” word representations, which are known as “word embeddings.”

We’ll use the sentiment-labeled sentences dataset from the UCI repository, consisting of 1,500 positive-sentiment and 1,500 negative sentiment sentences from Amazon. All the steps are detailed in the Word2Vec.

10- List the steps for converting training and test data into a format suitable for the neural network. ?

1. Tokenize the texts and convert them into word index vectors.
2. Pad the text sequences so that all text vectors are of the same length.
3. Map every word index to an embedding vector.
4. Use the output from Step 3 as the input to a neural network architecture

11- Which technique is better for text classification CNN or LSTM and why ?

☑ when the size of the data set is large or the sentences are long, it is preferable to use the LSTM.

☑ LSTMs and other variants of RNNs in general have become the go-to way of doing neural language modeling. This is primarily because language is sequential in nature and RNNs are specialized in working with sequential data.

☑ LSTMs are more powerful in utilizing the sequential nature of text, they’re much more data hungry as compared to CNNs.

12- How text classification models can be interpreted ?

☑ As ML models started getting deployed in real-world applications, interest in the direction of model interpretability grew. Recent research resulted in usable tools for interpreting model predictions (especially for classification). Lime is one such tool that attempts to interpret a black-box classification model by approximating it with a linear model locally around a given training instance.

13- How to solve no training and less training data problems?

☑ **No Training Data** → The first step in such a scenario is creating an annotated dataset

☑ **Less Training Data** → One approach to address such problems is active learning, approach for domain adaptation

14- Give some options to explore when no labels exist for a dataset.?

☑ Use existing APIs or libraries

☑ Use public datasets

☑ Utilize weak supervision

☑ Active learning

☑ Learning from implicit and explicit feedback

15- Describe the pipeline for building a classifier when there is no training data.?

READ and Understand ONLY

We start with no labeled data and use either a public API or a model cre- ated with a public dataset or weak supervision as the first baseline model. Once we put this model to production, we'll get explicit and implicit signals on where it's working or failing. We use this information to refine our model and active learning to select the best set of instances that need to be labeled. Over time, as we collect more data, we can build more sophisticated and deeper models.

Figure is important to draw in Exam !!

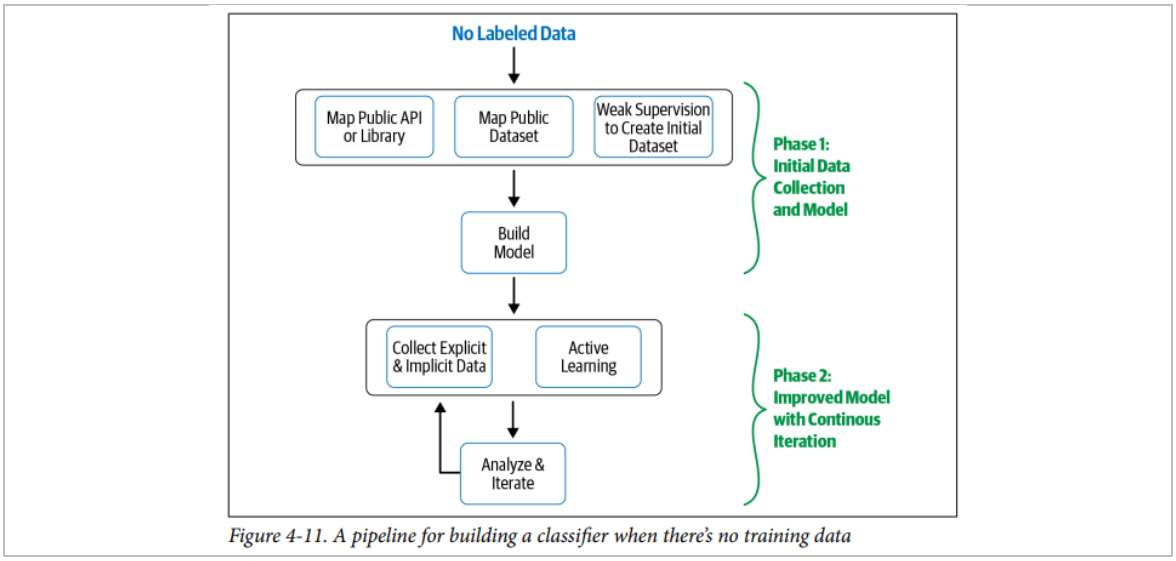

Figure 4-11. A pipeline for building a classifier when there's no training data