Mosaic ML Exercise

A classification problem to categorize news text into predefined groups.

Data:

Download the newsgroup data set from http://qwone.com/~jason/20Newsgroups/

Tasks:

Write your analysis, design and solution to the following tasks:

1. Classify the data using 4 traditional machine learning approaches (pick the models you think are helpful)
2. Classify the data using a neural network model.
3. Define, evaluate and compare the performance of the above models for doing this text classification.
4. What if the text is unlabeled short sentence, would you still consider to use the above models and why?
5. What would be your solution to classify any new unlabeled text?

Notes:

- Useful link and reference: https://docs.cloud.databricks.com/docs/latest/sample_applications/07%20Sample%20ML/MLPipeline%20Newsgroup%20Dataset.html
- Please finish the exercise independently within 5 days.
- Submit your exercise (email to cheng.he@saymosaic.com) in a pdf or slides with necessary answers, graph and code.