Ytremlet exercise

We prepared this “challenge” ahead of the meeting. The main interest is to see how you approach the problem and having a dialogue around the problem.

The data set is a drug review data set. The reviews are grouped into reports and ratings on the three aspects benefits, side effects and overall comment.

Data: Attached as two files (test_data.tsv, training_data.tsv)

Task: Predict the overall review, side effect and effectiveness rating based on the patient text reviews. Main focus should be on predicting the overall review rating based on the overall comment. The data is already split into training and test data.

Input:
Info: Drug name, Condition
Rating: overall, side effect and effectiveness ratings
Text: overall, side effect and effectiveness reviews

Output: Build a model based on NLP (Natural Language Processing) / Machine Learning techniques to predict the ratings in the test data and find meaningful ways of representing the data and results.

Comments: Usually when we get a reporting task, neither we or the people who requests the report knows exactly what we will find inside. Sometimes we know what we want, and formulate the questions so that we get our point through. Sometimes people have a query (e.g. how many of the drugs are recommended by the patients). Finally there is the completely open ended investigations, where we try to find patterns we did not know about before. The latest kind of report is usually the most valuable ones, business wise. It adds insight and learnings, while the others are more about driving an agenda.