Computational learning and data mining course completion project

The project deals with the classification of texts for medical communities.

The database was built by me and is the result of scraping on a health-related social network called "like me". It is a network made up of 41 communities dealing with various medical issues. The goal of the project was to receive posts and responses from the website and classify them using different types of models learned throughout the semester.
The special idea is to use the text itself as features rather than Metadata, and it requires trial and error to choose the optimal number of features for each classifier. This is chosen by calculating the tf-idf index for all the words in the corpus and choosing the X words with the highest ranking.
Of course, parameter tuning was also performed to optimize the models.

In terms of the project structure-

- In the main folder, 5 notebooks are the entire project in full, the files that can be re-run and get the results shown.
- In the data folder, there is the database file of the entire site as well as a stopwords file in Hebrew.
- The notebooks for the presentation folder contain the same notebooks, but for convenience, I deleted many of the outputs there and sometimes some of the code cells, to present in class concisely and focus on the unique conclusions for each notebook. There is a chance that it won't run because I deleted code cells and maybe there were settings there that disappeared, so to get an impression of the project and its conclusions this is the preferred folder, and this is also what I presented in class and you can see it in the recording. But to run and restore you need the original notebooks.

All notebooks are based on the same methodology - comparing all the classifiers, and the number of features, and optimizing the parameters. They only change in terms of the part of the dataset they check. Below is the difference:

1) The first notebook contains a comparison of all methods about 2 small and balanced medical communities. Seemingly the simplest task in the project.
2) The second notebook contains a comparison of 2 small but unbalanced communities (one more frequent.(
3) The third author tries to adapt the results from author 2 by balancing the communities with the help of oversampling.

4) The fourth notebook tries to adapt the results from notebook 2 by balancing the communities with the help of undersampling.
5) The second notebook contains a comparison of 6 larger and unbalanced communities - the most difficult task in the project.

At the end of each notebook, there is a summary and conclusions of the process of running in the notebook, and at the end of notebook 5 a summary and conclusions of the entire project.