Machine Learning project

Problem:
The dataset consists of the 25 top news headlines from reddit.com/r/worldnews and the stock data of the Dow Jones of every day for 8 years long. A day is labeled 0 if the Dow Jones decreased over the day, and labeled 1 if the Dow Jones increased over the day.

For the news headlines, we have 8 years of data, so 2920 days. Every day is a training example and consists of 25 headlines; you could say that the headlines are the features, but it is more accurate to say that the words used in the headlines form a set of features.

As mentioned above, all the days are labeled so we have 2920 y-values, either 0 or 1. Thus the problem is a classification-problem, and therefore, there are several algorithms applicable to this problem, of which some will be addressed below.
Additionally, the dataset consists of the actual stock data, if we have the time, it would also be interesting to see how much the Dow Jones changed, instead of just looking at the qualitative data.

Libraries to use
- SkLearn - to use classifiers on our data set, we need the SkLearn library.
- Sklearn Feature Extraction (Countvectorizer) - used to make vectors out of the words used in the headlines

Classification algorithms (classifiers)
- SVM
- Logistic Regression
(Are these two too similar? Should we maybe choose a different set of classifiers?)

Evaluation of classifiers
We should divide the dataset in a training set, cross-validation set, and a test set. After training different classifiers we can test them on the CV-set and test set and see if they can accurately predict whether the Dow Jones increases or decreases. The classifiers with the highest accuracy is the best classifier.

Algorithm
The most challenging part of this project is to transform the dataset from Kaggle into a dataset that we can use in our algorithm. This is done by converting the news headlines into vectors with words. If every day is a feature vector with words as components, we have a dataset that we can use.
This is where the countvectorizer comes into play. Another challenge in this process is making sure that different variants of words are considered as the same word. There should be no distinction between lower- and uppercase letters, singular and plural and conjugations of verbs.
Headlines most of the time do not contain words such as ‘the’, ’a’ or other meaningless words. If, however, these words are still in the headlines, we must find a way for our algorithm to ignore those words. (Does this matter? You could argue that the word ‘a’ is used in a decreasing DJ-day just as much as in an increasing DJ-day, and that is would therefore not influence the predictions of our algorithm.)

End result
If our algorithm is working correctly, we should be able to feed it with the 25 top headlines of any given day, and it should predict whether the Dow Jones will increase or decrease over the day.

Progress
This project-update-report is a good representation of our progress. We haven't had the time yet to really start working on the project e.g. actual programming or writing.