**Summary – Dinica Mircea**

The paper describes the development of a binary classifier for depression. Dataset used has 7650 entries which are labeled ‘0’ for no depression and 1 otherwise. The pre-processing tool used is named LIWC, which is based on a dictionary with categories of words as keys and the words which are part of that specific category as values. The classifier used was Random Forrest due to the best performance in comparison with multiple other classifiers found in a study on an input similar and output with ours. It was also the best overall across the rest of the datasets used in the study. For evaluation 4 metrics were used: classification metrics(accuracy, precision, recall, f1-score), confusion matrix. ROC-curve and feature importance.

For the first experiment the dataset was divided in a 75/25 train/test split, ensuring an equal distribution of positive and negative cases, and the given hyperparamaters from sklearn for RandomForrestClassifier were used. The model achieved 96% accuracy, having more false negatives than false positives, 67 compared to 6. In the second experiments all possible combinations for ranges found in a study for the hyperparameters were tested. The best possible combination achieved 97% accuracy, now having less false negatives, exactly 47. Also the top 10 most important features were more dispersed, word count and word per sentence having less of an importance.

Then the development of the classifier for Romanian language is detailed. The original dataset was translated using a python library named ‘googletrans’, which uses Google Translate API and is free to use. However the developers mention in the documentation that it is not the same as using the original API. The pre-processing tool LIWC latest available iteration for Romanian is LIWC-2015, which is less accurate and has some modifications in the categories compared to LIWC-22, which is only available in English. Hyperparameters were kept the same as the ones found the best in the second experiment for English. This classifier achieved only 92% accuracy, having 123 false positives, almost double compared to the English model. Also now the feature ‘word count’ had a bigger importance, which showed us in the English experiments that it is not ideal. The poor performance comes from the translation tool used and the decrease in version for LIWC.

Even with the chosen metrics, a professional psychologist would be needed to give a human-based assessment of the model. For future improvements a different, more accurate translation tool should be used and in order to use the model in business, the pre-processing method must be changed, because the LIWC documentation says that the tool is only intended for academic use.