CSC 480/680 – Assignment 2

Individual Assignment

Due on October 28, 2020

The pedagogical purpose of this assignment is to help you develop the following skills:

- Work with text data.
- Prepare raw text data for machine learning tasks using the Natural Language Tool Kit (NLTK).
- Experiment with new classifiers in addition to decision trees and multi-layer perceptrons: Naïve Bayes, k-Nearest Neighbors, Support Vector Machines, Bagging, and Boosting
- Experiment with unsupervised learning and more specifically clustering: k-means, agglomerative (single-link, Group average), self-organizing maps.
- Experiment with the UMAP visualization tool.

Note: this assignment is linked to an ongoing project that I am conducting with Professor Boukouvalas (Math/Stats) and a student. If you enjoy this work and are looking for a project topic for the course, you may want to explore a theme linked to this project in consultation with Prof. Boukouvalas and myself. This can also be the basis for an extended project outside the course. (E.g., an independent study or internship).

You will be provided with two data sets:

- *Genuine:* a data set containing genuine tweets with no misinformation about Covid-19
- *Mixed_Misinformed:* a data set containing tweets that have been filtered by keyword searches and that contain some tweets with misinformation about Covid-19 in them, but others about such misinformation.

The practical purpose of this assignment is to see if you can reduce the *Mixed_Misinformed* data set to a set, *Misinformed*, that includes only tweets containing misinformation. One idea for doing that could be the following iteration:

1. Prepare the data for classification and clustering (see below).
2. Set S = Genuine vs Mixed_Misinformed
3. Use a series of classifiers (those listed in bullet point 3 above including Decision Trees and MLP) to discriminate between Genuine and Mixed_Misinformed tweets. From your results, select the classifier you "trust" most, TClass, for its ability to classify this data.
4. Assuming that the tweets in *{Mixed_Misinformed – Misinformed}* contain tweets that resemble the Genuine tweets more than the Misinformed tweets, cluster the *Genuine +*

Mixed_Misinformed tweets using all the clustering approaches listed above (in bullet point 4) as well as UMAP (in bullet point 5).

5. Choose the clustering method you trust most or select to continue with UMAP. Call your chosen method TClus.
6. Use TClus on S and mark the tweets from *Mixed_Misinformed* that lay closest to the *Genuine* ones as *Probably_Genuine*. (Use your own strategy to select the *Probably_Genuine* tweets).
7. Modify S by transferring the Probably_Genuine data to the Genuine class (or simply eliminating the Probably_Genuine data altogether) and using the tweets that remain in Mixed_Misinformed as the Misinformed class.
8. Run TClass on S
9. If TClass' performance has significantly improved then go to 6.
10. Otherwise return the tweets contained in the Misinformed class.

Note: this approach relies on the hypothesis that the classification problem will be simplified if the data in Mixed_Misinformed that is not Misinformation gets transferred to the Genuine data set. This hypothesis may be wrong. We will see if it is when we compare the Misinformed data set you obtained with a gold standard set that was generated by two human subjects including a linguistic expert. (that will be done as a blind test: you will be asked to submit your Misinformed tweets.

Method for preparing text data:

The following tutorials can help you figure out how to turn tweets into vectors that can be used in your classifiers.

https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/

https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/

Note: for the purpose of this assignment, you can represent the tweets using tf-idf vectors. However, if you are interested in experimenting with word embeddings, please feel free to do so using Word2Vec, BERT or other pre-trained model such as one specifically created for Covid-related tweets (https://arxiv.org/abs/2005.07503). You can get up to 10 bonus points if you consider both tf-idf and the word embedding route.

Your report should be divided in two parts.

Part I is there to show that you ran all the algorithms that I asked you to run on the original data provided (prior to your reducing the Mixed_Misinformation set. Please, output:

- A table showing the results you obtained with all the classifiers (please list the precision, recall and F-Measure). Note: run 10-fold cross-validation and list the results obtained at each fold. Treating each fold as a different domain, use Friedman's Test followed by Nemenyi's test to assess whether the difference observed between the classifiers are statistically significant.

- A table showing the results you obtained with all the clustering approaches (please list the silhouette coefficient, the Calinsky-Harabasz Index and the Davies-Bouldin Index).
- A UMAP plot.

Please discuss all the results you obtain.

Part II will describe the strategy you devised to eliminate data from Mixed_Misinformed. You should also explain whether you eliminated the Probably_Genuine instances or added them to the Genuine data set. It would be great if you could plot the UMAP results at each iteration to show the evolution of your Mixed_Misinformed data set. Please conclude Part II by commenting on whether you believe that the hypothesis on which this assignment is based was correct or not.