Sentiment Analysis

Arda Derbent, Anton Yahorau
Group 21

April 2023

1 Preliminary Documentation

- A short description of the algorithms that will be used, along with some examples.

  Algorithms: Logistic Regression, Naive Bayes and Support Vector Machines will be used. These are classification algorithms that scikit-learn provides

  Feature extraction methods: Bag of Words, tf–idf and Hashing vectorizer will be used. These are feature extraction methods necessary for sentiment analysis

- Selection and description of the datasets.

  Twitter US Airline Sentiment. This dataset contains twitter data that analyzes how travelers in February 2015 expressed their feelings about Airlines

  https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

- General plan of tests/experiments.

  Only using Sentiment score and the relevant text data columns

  With and Without parameter tuning each algorithm will be tested using different algorithms and feature extraction methods

  Total of 18 combinations of 3 feature extraction methods will be obtained with different accuracies and cross validation scores

  The solutions obtained from previous experiments will be combined and a more detailed analysis/comparison will be made

  These comparisons will be made using bar graphs using their accuracies, cross validations and AUC scores

- Methods of result visualization.

  We plan on using bar graphs with the accuracy outputs of various classification algorithms and their cross validation scores. Also AUC scores on bar graphs

- Definition of quality measures that will be used. Cross validation scores and AUC scores will be used.

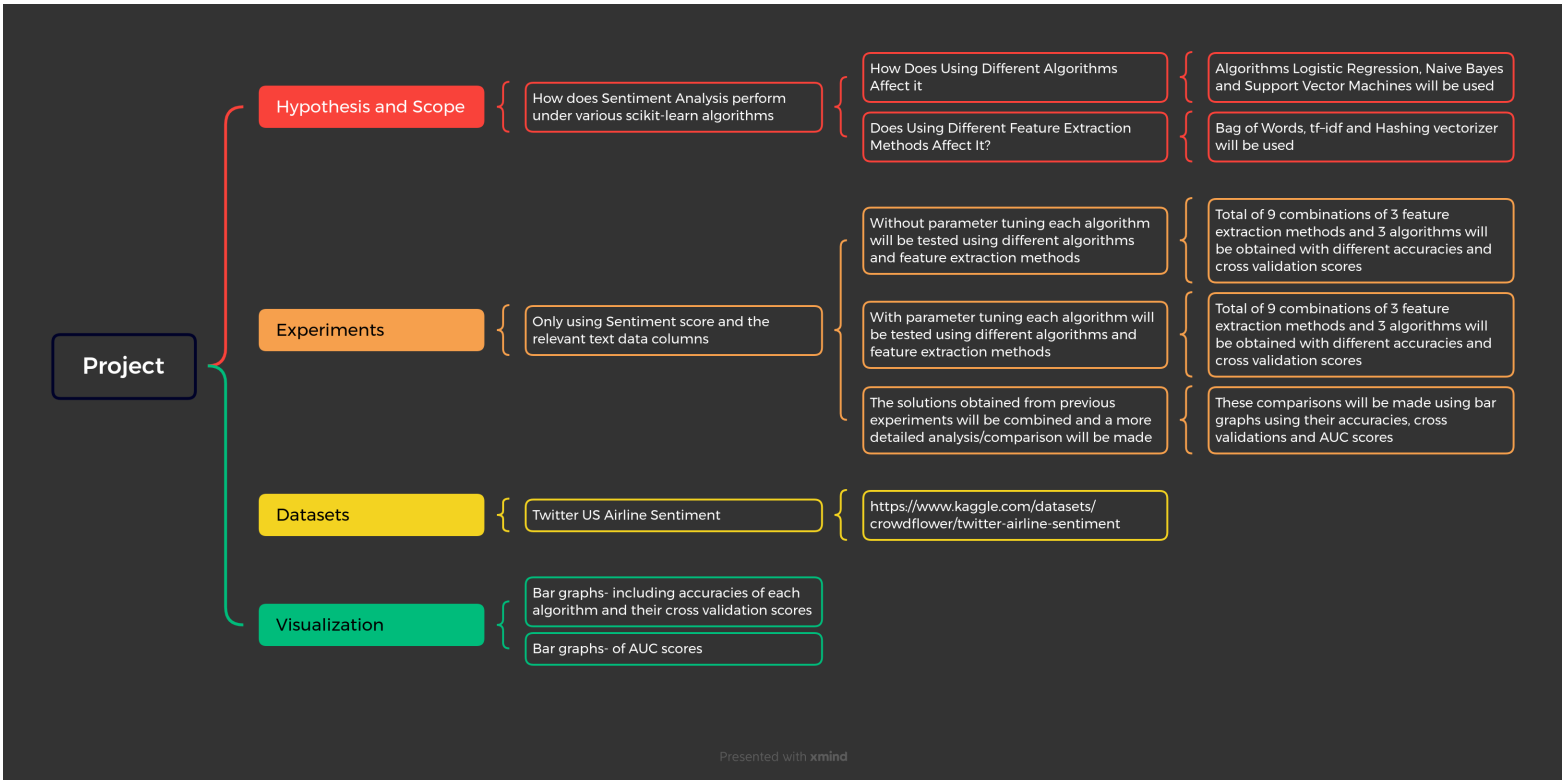

Figure 1: Roadmap for replicating (2).