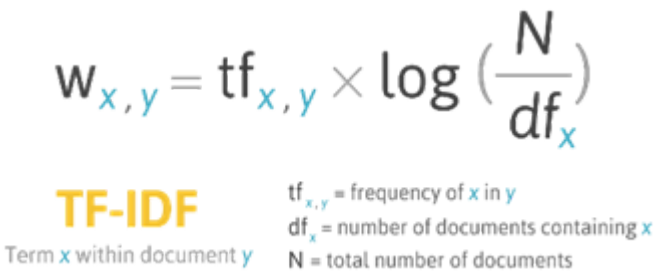
University of Crete
Department of Physics

Subject:
Detecting Fake News using machine learning techniques

Team Members:
Ειρήνη Κωτσίδου ph4706
Ανδρέας Φωτιάδης ph5368
Κυριακή Μπιμπίρη ph4725

Abstract

The purpose of this project is to detect whether a news article is a fake or real source of information by using six different machine learning classifiers. The shape of the training dataset is (20800,5) and after essential processing we reduce the shape to (20387,5). When the preprocessing is done, the dataset is split into a training set, which includes 80% of the data, and the test set, which includes 20% of the data so we can proceed to machine learning techniques in order to identify the reliability of the news. The results point out that the accuracy of the models is quite satisfactory with XGBoost giving the accuracy score of 0.97.



The dataset we used can be found on the Kaggle website and contains the following attributes: *id*: unique id for each article, *title*: the title of a news article, *author*: the author of the news article, *text*: the text of the article; could be incomplete and the *label*: a label that marks the article as a potentially unreliable. In specific, the labels consist of 1 and 0, for fake and real news respectively. For the purpose of the project, we used Python 3.10.

At the beginning of the given Jupyter Notebook we imported all the important libraries. After that we uploaded the training dataset by using pandas. The shape of the dataset is read to be (20800,5). As mentioned before, the text of some articles could be incomplete or we could find some duplicates. By checking for missing values and for duplicates the shape of the dataset was reduced to (20387,5).

PRE – PROCESSING

At the beginning of the pre – processing, we used a text normalization technique, which is called stemming, in the field of Natural Language Processing to prepare the text. This process reduces inflection in words to their root forms. Stopwords (words that a search engine has been programmed to ignore) were removed. We defined X, Y values: X as the title, text & author and Y as the label. For X, Y values, we used TF – IDF (term frequency – inverse document frequency) , a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

Last step of the pre – processing was to convert text into numbers. Dataset was randomly split into a training set (80% of the data) and a test set (20% of the data).

METHODS

The machine learning classifiers we used are based on the following methods.

1. *Logistic regression* is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).
2. *A random forest* is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the parameter if bootstrap = True (default), otherwise the whole dataset is used to build each tree.
3. *XGBoost* is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting. It provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way. The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.
4. *Support vector machines (SVMs)* are a set of supervised learning methods used for classification, regression and outliers' detection. The advantages of support vector machines are: Effective in high dimensional spaces. Still effective in cases where number of dimensions is greater than the number of samples. Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient. Versatile: different Kernal functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels. The disadvantages of support vector machines include: If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernal functions and regularization term is crucial. SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.
5. *Decision tree learning or induction of decision trees* is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification – trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression – trees. Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.
6. *The passive-aggressive* algorithms are a family of algorithms for large-scale learning. They are similar to the Perceptron in that they do not require a learning rate. However, contrary to the Perceptron, they include a regularization parameter C.

For each of these methods, we calculated the accuracy on the test and on the train set. We also found the predicted Y value and built a confusion matrix to observe better the extracted results. We defined in order to print the accuracy score (number of correct predictions), the precision score (tp: true positive, fp: false positive), the recall score (tp: true positive, fn: false negative) and f1 score (the weighted average of precision and recall score). By plotting some of the results we were able to compare the accuracy of the models and the time needed for the process.

EVALUATION

The last step of the project was to check whether the machine learning methods we used are appropriate for detecting fake news and how well they perform. In this part, the test set was used for the evaluation. We randomly defined a value as one of the values and we used predictors in order to decide the reliability of the news article.

The results indicate that XGBoost classifier operates the best and that machine learning can be used for this cause. We also performed the same procedure for a dataset with shape of (6335,4) and the results were not as good as in the bigger dataset. This leads us to the conclusion that the model can be trained to learn better with more data.

References

1. (Kaggle, n.d.)
2. (Scikit-learn - machine learning in python, n.d.)
3. (Python, n.d.)