Tajudeen Abdulazeez: IST 738 Text Mining

Fedpapers Dispute: A Kappa Measure Of The Agreement Between Supervised And Unsupervised Machine Learning

Author*
Syracuse University, School of Information Studies, Syracuse,NY

The main objective of this paper is to examine the agreement between supervised machine learning and unsupervised machine learning in predicting the real author of the disputed federalist papers.The following unsupervised techniques; KMeans cluster, hierarchical cluster and agglomerating clustering, were used to classified the disputed papers. The algorithms used for the supervised machine learning are support vector machine and Multinomial Naive Bayes. Kappa measure is used to measure the agreement between the result from supervised machine learning and unsupervised machine learning.

Keywords: Fed-papers,Kappa Measure,Machine Learning, Text Mining

I. INTRODUCTION

The Federalist Papers were a series of eighty-five essays urging the citizens of New York to ratify the new United States Constitution. Written by Alexander Hamilton, James Madison, and John Jay, the essays originally appeared anonymously in New York newspapers in 1787 and 1788 under the pen name "Publius." A bound edition of the essays was first published in 1788, but it was not until the 1818 edition published by the printer Jacob Gideon that the authors of each essay were identified by name. The Federalist Papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution (quote from library of congress).This make it very interesting resolving the authorship dispute as both Hamilton and Madison claims the authorship. An author is the person who originate or gave existence to anything and whose determined the responsibility for what it was created.There are about 11 papers which original authors are disputed. Different approach of machine learning has been used to try resolving the dispute. The aim of this experiment to use kappa measure to compare the results obtained from both supervised machine learning and unsupervised machine learning to ascertain there there agreement which will increase the level of confidence of the actual authors of the disputed essays.

II. ANALYSIS AND MODEL

The analysis is carried out using python programming language with the following packages;pandas, numpy, wordcloud, KMeans, and SKlearn. The document were pre-processed and vectorized using Countvectorizer and TF*IDF vectorizer before using the clustered algorithm to predict who wrote the disputed essays based on the cluster centroid.

There are 74 essays with identified authors: 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay, the remaining 11 are either authored by Hamilton or Madison. The features are a set of “function words”. The feature value is the percentage of the word occurrence in the essay.

III. ANALYSIS

The text document originally downloaded from the library of congress is converted to a document term matrix where the feature values is the percentage of the word occurrence in the essay. The document label is ignored for the clustering analysis.

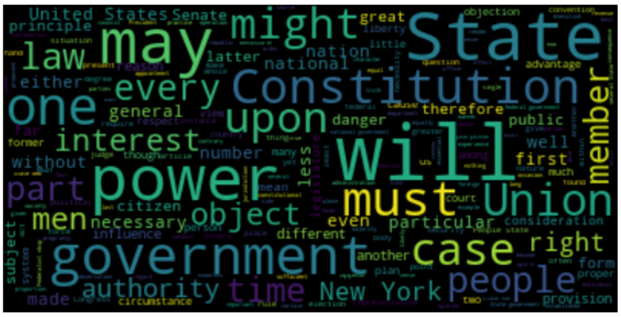


Figure 1. Wordcloud from the corpus.

The most occurrence words using the word cloud from figure 1 are Constitution, State, Power , government, Union, law, interest, etc. which are inline with the context of the documents.

* Correspondence email address: toabdula@syr.edu

Sklearn CountVectorizer is used to transform the datasets into a vector that algorithms can work with.

Different ngram number combination were tried to evaluate the performance of model and also different type of vectorization process. The vectorizer with the best performance is used for the final model.

Let's get a little dive into what vectorization process means and the different types of vectorization. Vectorization is the process of transforming text document in a way computer can understand. The only way a computer can understand text is to represent with number which is call text vectorization. This is an important step in data mining. Each word represented by a number is called a token. Tokenization rules can vary depending on the business problems and how grouping of words can have significant impact in identifying trend in text document. The tokens can be represented with n-grams or bag of words. Bag of word tokens ignore the context of words while n-grams while n-grams capture the local content of words, the representation can be unigrams, token of individual words, bi-grams, token of two words, tri-grams, token of three words and so on. There are four major different type of converting text to a vector for computer to understand.

1. Boolean Vectorizer 2. Count or Term Frequency Vectorizer 3. Normalized term Frequency 4. TF*IDF Vectorizer

Boolean is a vectorization method where each token is represented, either if it is present in a document or not. If a token is present, it is represented by 1 and 0 otherwise. Term Frequency is a vectorization method where each token is represented by the number of occurrences in a document. If you a document and a token appeared in the document 5 times, the token or word will be represented by 5. Normalize term frequency is a method of vectorization where word frequency is normalized by the document length. If you have a document with a length of 1000, and a word appeared in the document for 20 times, the normalized term frequency will be 20/1000. Tf*idf is a concept borrowed from information retriever, and it's a blind weighting strategy for text classification. in this method, the term frequency is multiplying by the inverse document frequency. This is aimed to penalize the common word across documents.

IV. MODEL

Support Vector Machine : An SVM is a supervised machine learning technique that uses support vectors in classification. The support vectors define the hyperplane which are the most important part of training of an SVM. Any line in dimension D can be represented as w T x + b = 0. w is a vector of coefficients, x is the vector of variables, b is the translation (can be thought of in 2D as the y intercept) as illustrated in figure 2.

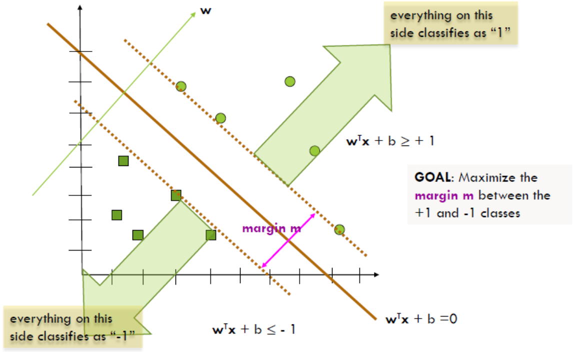

w
everything on this side classifies as "1"
$w^T x + b \geq + 1$
GOAL: Maximize the margin m between the +1 and -1 classes
margin m
everything on this side classifies as "-1"
$w^T x + b \leq - 1$
$w^T x + b = 0$

Figure 2. Setting up the SVM Problem. The goal is to maximize the margin m between the +1 and -1 classes.

SVM uses different "kernels " options such as linear, gamma, sigmoid, and Gaussian. SVM can perform feature transformation into higher (or infinite Hilbert Space) dimensional space so that input vectors (x) are separable by hyperplane (high dimensional plane). It can be used for regression, classification and outlier detection and below are the summary of advantages of SVM:

1. Effective in high dimensional (D) space 2. Effective when dimension is greater than the number of samples 3. It uses subset of the training data, just the support vectors 4. Allows or the use of kernel functions

While some of the disadvantages are summarized below: 1. If the features is greater than the number of samples, there will be poor performance 2. No probability estimates- these must be done by hand using cross validation. SVM did not required any special form of vectorization process. It will work fine with any form of text transformation compared to MNB that required a count vectorizer as input.

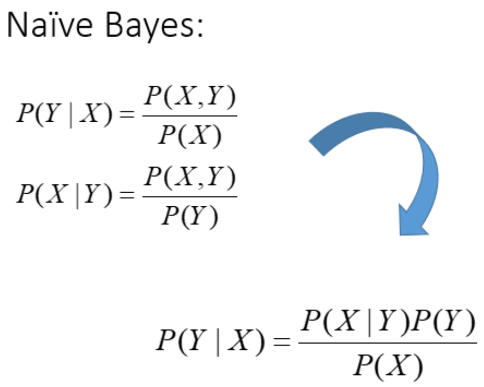
MNB (Multinomial Naïve Bayes): MNB used the Bayes theorem assumption in classification problem. It assumed that all attributes are independent which make it suitable for text classification.

Naïve Bayes is robust to isolated noise, it handle missing values by ignoring instances during probability estimate calculations. MNB required a count vectorizer as input for optimal performance.

K-Means Clustering: K-means assigned a cluster centroid to each of the cluster point. Each of the point is assigned to the cluster with the closest centroid. The number of K (cluster) must be specified and the objective is to minimize the sum of distances of the points to their respective centroid.

Hierarchical Clustering: The two main type of hierarchical clustering are also use in this project.

1. Agglomerative which start with the points as in-



Naïve Bayes:

$$P(Y\mid X)=\frac{P(X,Y)}{P(X)}$$

$$P(X\mid Y)=\frac{P(X,Y)}{P(Y)}$$

$$P(Y\mid X)=\frac{P(X\mid Y)P(Y)}{P(X)}$$

Figure 3. Naïve Bayes

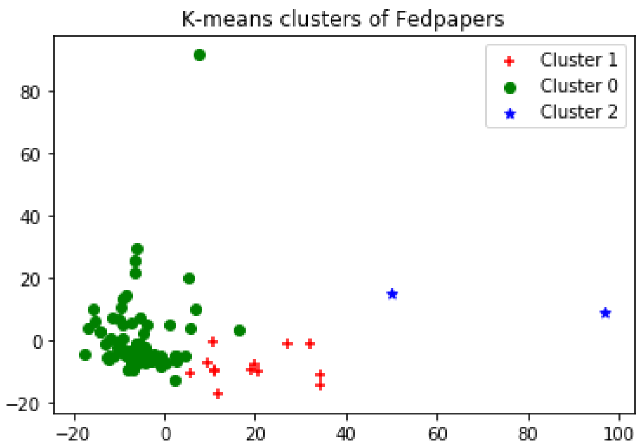


Figure 4. K-Means Cluster, k=3

dividual clusters, at each step, merge the closest pair of clusters until only one cluster (or K clusters) left.

2. Divisive start with one, all-inclusive cluster. At each step, split a cluster contains a point (or there k clusters) which can be visualized using the dendogram.

Kappa: Kappa values $<$ 0 as indicating no agreement and 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement. (Landis and Koch)

V. RESULTS

The kappa score of the results obtained from unsupervised machine learning (Hierarchical cluster) and Supervised machine learning (Multinomial Naive Bayes) is 0.62.

Multinomial Naive Bayes is used to make prediction based on its performance in model evaluation, compared to support vector machine while hierarchical clustered is used for the unsupervised because the prediction can easily be visualized and explained using the dendogram

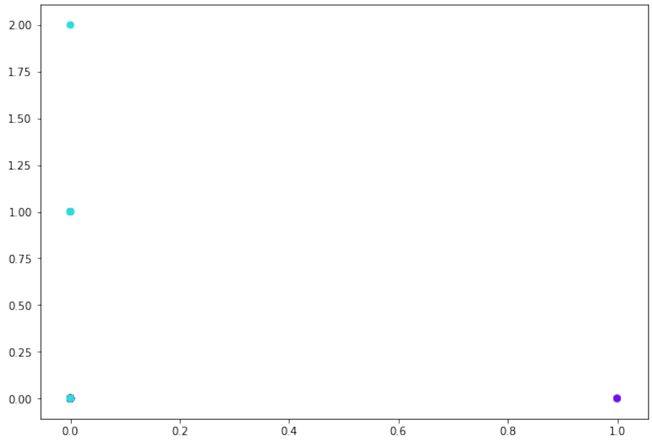


Figure 5. Agglomerative Cluster

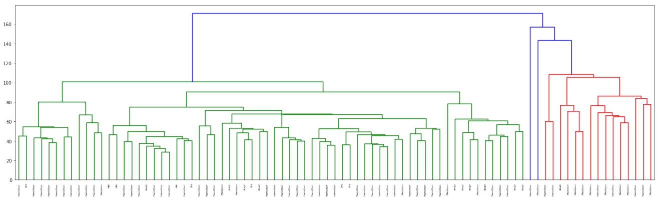


Figure 6. Divisive Cluster

VI. CONCLUSION

Supervised (MNB) and Unsupervised(Hierarchical Cluster),Substantially agree in predicting the author of the disputed essays using Kappa measure.

The majority of the disputed paper are written by Madison based on the two approached of machine learning techniques.

VII. REFERENCES

https://drive.google.com/drive/u/1/folders/1XjJeYe1ztaw8U