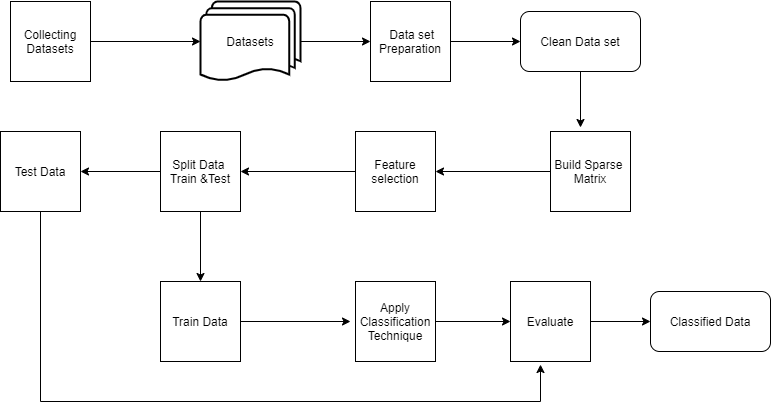
Research topic: Classification (News classifier)

# 1. Introduction:

Every day millions and millions of news are being published around the world, each one of them is telling an information, giving hope, expressing people opinion or telling a story of someone ‘s victory each one of them is considered valuable to a user with particular interest. Day after day this data are being available online in text form and classifying them became more difficult but with ML (Machine Learning) and the available new tools they provide very useful classifications to these texts. But the steps of the process require some preparing steps which is essential for the process and for enhancing the results in this document I will explain the steps of my news classifier, including preprocessing steps, the classification models I applied to the dataset and accuracy of each one of them, no one will be accurate 100%; but as it is known in ML “All models are wrong but some are useful”.

# 2. Classification workflow

The process of classification is difficult and required some steps, including collecting dataset, preparing it, build the sparse matrix, split the data into test and train and apply different models.



Process \_Workflow Fig (1)

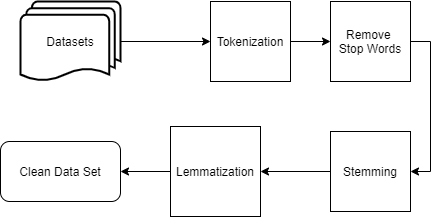
**Collecting Dataset:**

News Data is now available in many sources like online Newspaper, Magazines, radio, TV, Press or WWW(world wide web) but also indifferent forms and extensions like; DOC, HTML, XMl, pdf or any other format.

In this Classifier two Datasets are being used small dataset about 1400 article from BBC news and big data about 120000 articles from different resources

**Data Preprocessing:**

As data has been collected from different resources with different format it has 2 main columns the category column that need to be predicted, and don’t need any preprocessing steps. But the article column that is the one need all preprocessing steps to be applied to. each article contain words that has no importance to the article context and mane words that has the same meaning will only increase the number of features in the matrix for that cleaning the Dataset saves time ,memory and improve the accuracy . the process has number of steps each one target particular domain with cleaning data.



Preprocessing Fig (2)

**Data tokenization:**

It is the process of splitting data into words each is cindered a string and it is a significant part as its output is the next process input.

**Remove Stop Words:**

Each language has its own stop word that doesn’t add meaning to the sentence like delimiters, pronouns, and Prepositions. This process is applied her by downloading the language stop words and remove it from tokenized words

**Data Stemming:**

After removal of stop words a stemmer is used to remove suffixes, prefixes and punctuations from the words returning many words to the same original word. In this process a Porter Stemmer is used due to its simplicity and precision that is why it is the most used stemmer.

**Lemmatization:**

Stemming is not always enough to get the root of the word for some irregular verbs they will not be affected, in different cases it ruined the meaning of the word unlike lemmatization which return the word to its original root. With a very popular example to demonstrate the differ between stemmer and lemmatize

Word: was stem: wa lemma: be

The stemmer considered (s) in (was) as Suffix and remove it while lemmatization return the word to its original root (be) that show its importance in cleaning process.

**Build sparse matrix:**

Sparse matrix is a matrix contains a small number of non-zero elements. It is very useful in solving and representing many algorithms and is used in many applications.

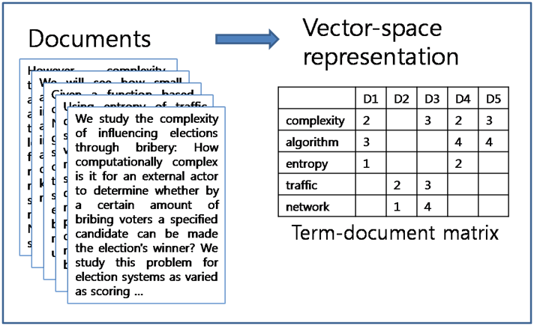


Figure 3- source [4]

In this classifier a TfidfVectorizer is used to build the matrix which use the bag of words generated from preprocessing phase but it does not only count frequency, it also gives weight to every word showing how significant it is as a feature. And it provides the feature to choose the number of ngram 1, 2 or 3 showing how correlation between words as a new features.

**Feature selection:**

Large number of data produces huge number of features, which if it wasn’t important it will cost time and memory and sometimes won’t reach the desired accuracy.

In this classifier chi-square is used to reduce the dimensionality of the matrix making the model with much less number of features without affecting the accuracy or make it slightly better.

**Splitting Dataset:**

Each model in ML need a large number of data to train the more is better but it isn’t right for the model to be tested over the same data it was trained so the data has to be split to test part and train part sometimes it would be split to a third part for evaluation to avoid over fitting but in this case the data will be split into training set (80%) and test set (20%).

**Classification techniques applied to the data and evaluation:**

|  |  |
| --- | --- |
| Classification Technique | intuition |
| LSTM | Long short-term memory (LSTM) it is a deep learning algorithm that is a part of RNN but has memory that can process a sequence of data not only a single point. |
| SVM | Support vector machine: is a machine learning algorithm that is used for data classification and regression analysis. It is a supervised learning method that produces a map of the sorted data with the margins between the two as far apart as possible. SVMs are used in text categorization, image classification, and handwriting recognition |
| Logistic Regression | Logistic regression is a type of statistical analysis which is used for predicting the outcome of a dependent variable according to prior observations. It is a very common machine learning algorithm that has many applications for example it can predict the results of an election based on previous election results, economical and political data. |
| naive Bayes | A naive Bayes is an algorithm that uses Bayes' theorem for classifying objects, it makes strong, or naïve assumptions, without depending on the attributes between data points. It has many applications in text analysis, medical diagnosis and filtering spams. |
| Decision Tree | A decision tree is a graphical representation used for complex branching situations. It is a predictive model that produces generalized output based on Boolean series of specific situations.  Each node represents a decision point. Rules in the decision tree are represented by paths between the root to a selected action. |
| random forest | A random forest is a machine learning technique where large numbers of random decision trees are developed for analyzing collection of variables. And it helps in enhancing the accuracy of analyzing complicated data. |
| k-nearest-neighbor | A k-nearest-neighbor algorithm is an approach to data classification which estimates the likelihood of a data point to be a member of a group or the nearest group that the point could belong to. As it is a type of lazy learner techniques. |

1. **Dataset 1 (4 categories :world/sports/business /since)(120000 article)**

|  |  |  |
| --- | --- | --- |
| Classification Technique | Accuracy | Confusion matrix |
| LSTM | 0.9395 |  |
| SVM | 0.90775 |  |
| Logistic Regression | 0.906458333333333 |  |
| Naive\_bayes | 0.808 |  |
| Random Forest Classifier | 0.8630416666666667 |  |
| Decision Tree | 0.811458333333333 |  |
| K Neighbors Classifier | 0.637291666666666 |  |

Comparison between dataset(1) accuracy

**2-Dataset 2 (3 categories: politics/sports/ business) (1438 article)**

|  |  |  |
| --- | --- | --- |
| Classification Technique | Accuracy | Confusion matrix |
| LSTM | 0.9957 |  |
| Logistic Regression | 0.982638888888888 |  |
| SVM | 0.979166666666666 |  |
| Random Forest Classifier | 0.96875 |  |
| Naive\_bayes | 0.951388888888888 |  |
| K Neighbors Classifier | 0.899305555555555 |  |
| Decision Tree : | 0.892361111111111 |  |

Comparison between dataset (2) accuracy

**Prediction:**

After running model for training 3 parts are saved as separate files on the disk which are (classifier model, TfidfVectorizer and chi square model).

Classifier model: for prediction, TfidfVectorizer: for transferring the article to single vector

And chi-square: for reducing the features of the sentence to the selected number.

Then in the prediction model the 3 files are uploaded and the input article pass the steps of cleaning its text, get transferred by TfidfVectorizer and reducing the features by chi-square and finally calling the model for predicting which class it belongs to.

**Conclusion:**

Building a classifier does not only depend on selecting a ML technique, it requires to apply every step carefully from collecting data and clean them well to feature selecting and then choose the right technique.

The different dataset shows that the number of classes affect the accuracy as the more classes is add the less accuracy would be achieved.

Not every ML technique get the same accuracy with all data; in this classifier logistic regression and SVM get the highest accuracy, although Logistic regression did not take as much time in learning and predicting as SVM. This does not prove other techniques is not good it is just they didn’t get the best accuracy in this case.

**References:**

[1] OSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 18, Issue 1, and Ver. III (Jan – Feb. 2016), PP 22-26, News Classification and Its Techniques: A Review Gurmeet Kaur, Karan Bajaj

[2] International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017 ISSN: 2231-2803 http://www.ijcttjournal.org Page 128 Supervised Machine Learning Algorithms: Classification and Comparison Osisanwo F.Y, Akinsola J.E.T., Awodele O, Hinmikaiye J. O., Olakanmi O, Akinjobi J.

[3] Aegean International Textile and Advanced Engineering Conference (AITAE 2018), Automated News Categorization using Machine Learning methods U Suleymanov1 and S Rustamov

# [4]Quora, What is a tf-idf vector,Jul 19 2019