



UNIVERSITÀ DI PISA

Artificial Intelligence and Data Engineering

Data Mining and Machine Learning

News Classifier

Project Documentation

AUTHOR:
Leonardo Bargiotti

Contents

1	Introduction	3
1.1	Goals	3
1.2	Initial Dataset	3
1.2.1	AG News	3
1.2.2	BBC News	3
2	Preprocessing	5
2.1	Removing Duplicates and Missing Values	5
2.2	Text Preprocess	5
2.2.1	Normalization	5
2.2.2	Stemming	6
2.2.3	Lemmatization	6
3	Classification	9
3.1	Vectorization	9
3.2	Classifiers	9
3.2.1	Hyperparameters	10
4	Results	12
4.1	AG News	12
4.1.1	Classification Report	12
4.1.2	Confusion Matrix	13
4.1.3	Area Under the Curve	13
4.1.4	Class Prediction Error	14
4.2	BBC News	14
4.2.1	Classification Report	14
4.2.2	Confusion Matrix	15
4.2.3	Area Under the Curve	15
4.2.4	Class Prediction Error	16
5	Application	17
5.1	How to Install	17
5.2	How to classify text	17
5.3	How to change dataset	18
5.4	Display statistics	19

List of Figures

1	Class Distribution AG News	5
2	Class Distribution BBC News	5
3	Wordcloud AG News	6
4	Wordcloud BBC News	7
5	Top 20 Words on AG News	7
6	Top 20 Words on BBC News	7
7	Before and After Preprocess	8
8	Confusion Matrix AG News	13
9	Area Under the Curve AG News	13
10	Class Prediction Error AG News	14
11	Confusion Matrix BBC News	15
12	Area Under the Curve BBC News	15
13	Class Prediction Error BBC News	16
14	Home Application	17
15	Configuration	18
16	Statistics	19

1 — Introduction

Unstructured data in the form of text: chats, emails, social media, survey responses is present everywhere today. Text can be a rich source of information, but it can be hard to extract insights from it. Text classification is one of the important task in supervised machine learning (ML). It is a process of assigning tags/categories to documents helping us to analyze automatically text in a cost-effective manner. It is one of the fundamental tasks in Natural Language Processing with broad applications such as sentiment-analysis, spam-detection, topic-labeling, intent-detection etc.

This project exploits text mining in order to automatically categorize news articles to their right topic (for example sport, business, world and sci/tech). This application could be used to classify text of other subject (not only news), depending on which dataset is given in input.¹

1.1 Goals

The aim of this paper is to explain the choices and the strategies adopted on the project and development of **News Classifier**. In order to accomplish it, the first step is perform preprocess for having a suitable dataset and then it is used several classifiers, to determinate which predicts the right class.

1.2 Initial Dataset

In order to realize this application are used two different dataset:

1. **AG News**²

2. **BBC News**³

1.2.1 AG News

The first dataset is composed by 120000 rows in training set and 7600 in testset. It is perfectly balanced in fact it has four class *World*, *Business*, *Sports* and *Sci/Tech* and each one have the same number of instances, in particular 30000 in training set and 1900 in test set. It has three columns: one relative to the class of the news, the second is the title of the news and the last one is the news. For this application the title is not useful. The column, relative to classes, contains numbers associate to the four classes *1,2,3,4* are associate respectively with *World*, *Sports*, *Business* and *Sci/Tech* class.

1.2.2 BBC News

In the second one is smaller than previous, it has only 2225 rows in dataset (it will be divide into training and test set using *train_test_split*⁴, default value of *test_size* is 0.3). It has five classes *Sport*, *Business*, *Tech*, *Politics* and *Entertainment* and they are not balanced, the class most frequent is Sport with 511 instances and the class minus frequent is Entertainment with 386. The are only two columns: one relative to the description of the news and the other to the corresponding class. The column, relative to classes, contains strings that they will be encode

¹Github repository for the project: <https://github.com/leobargiotti/News-Classifier>

²Link for AG News Dataset from Kaggle: <https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset>

³Link for BBC News Dataset from Kaggle: <https://www.kaggle.com/datasets/yufengdev/bbc-fulltext-and-category>

⁴Link for *train_test_split* method: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

to number using *LabelEncoder library*⁵. In both datasets there aren't missing value but there are 1185 and 99 duplicates respectively on the first and second one.

⁵Link for LabelEncoder library: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>

2 — Preprocessing

Before building models, is necessary preprocess dataset in order to remove first of all missing values, duplicates and then clean the text. The entire process is shown in this chapter.

2.1 Removing Duplicates and Missing Values

The first thing is remove duplicate rows and those contain missing values. In the images below is possible to see that after removing values the new datasets are balanced as originals and the number of elements of each class are quiet the same that previous.

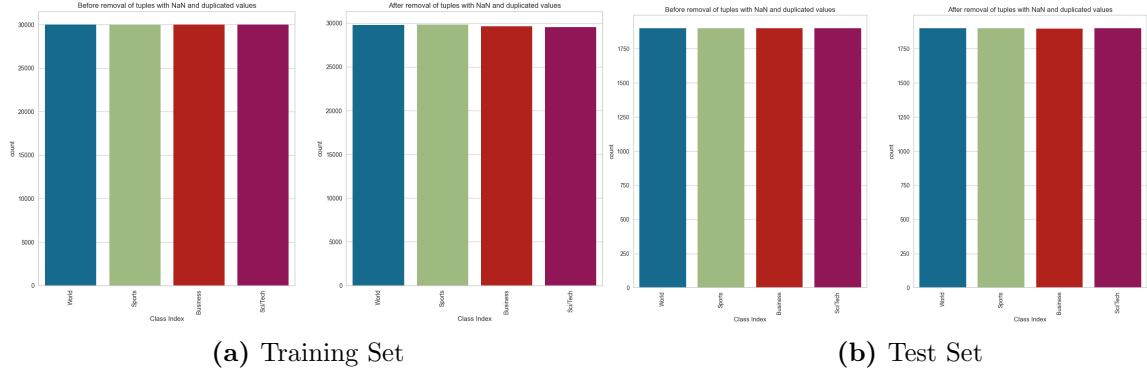


Figure 1: Class Distribution AG News

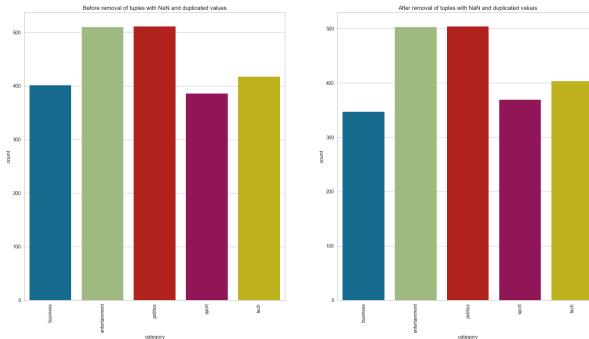


Figure 2: Class Distribution BBC News

2.2 Text Preprocess

In order to perform computational tasks on text, is need to convert the language of text into a language that the machine can understand. In particular text cleaning is composed by following steps:

- Normalization
- Stemming
- Lemmatization

2.2.1 Normalization

One of the key steps in processing language data is to remove noise so that the machine can more easily detect the patterns in the data. Text data contains a lot of noise, this takes the form of

special characters (such as URLs, HTML tags, diacritics, extra white spaces), punctuation and numbers. Additionally, it is also important to apply some attention, if text includes both upper case and lower case versions of the same words then the computer will see these as different entities. To avoid this problem is enough transform all words in text to lowercase. The python library used to implement this steps is *texthero*⁶. Another important phase is removing stop-words, it is list of generic words for example ‘*i*’, ‘*you*’, ‘*a*’, ‘*the*’, ‘*he*’, ‘*which*’ etc. for the English vocabulary. The list of stop-words used is the default in *nltk library*⁷. There are another feature only available for English text, is to write abbreviations in their long forms and slangs in to the correct form, using *contractions library*⁸.

2.2.2 Stemming

Stemming is the process of reducing words to their root form. For example, the words '*rain*', '*raining*' and '*rained*' have very similar, and in many cases, the same meaning. The process of stemming will reduce these to the root form of '*rain*'. This is a way to reduce noise and the dimensionality of data. To implement this process is used *texthero library*, used in the previous step.

2.2.3 Lemmatization

The goal of lemmatization is the same as for stemming, in that it aims to reduce words to their root form. However, stemming is known to be a fairly crude method of doing this. Lemmatization, on the other hand, is a tool that performs full morphological analysis to more accurately find the root. To implement this *simplemma library*⁹.

After apply this process to the original datasets these are the results:

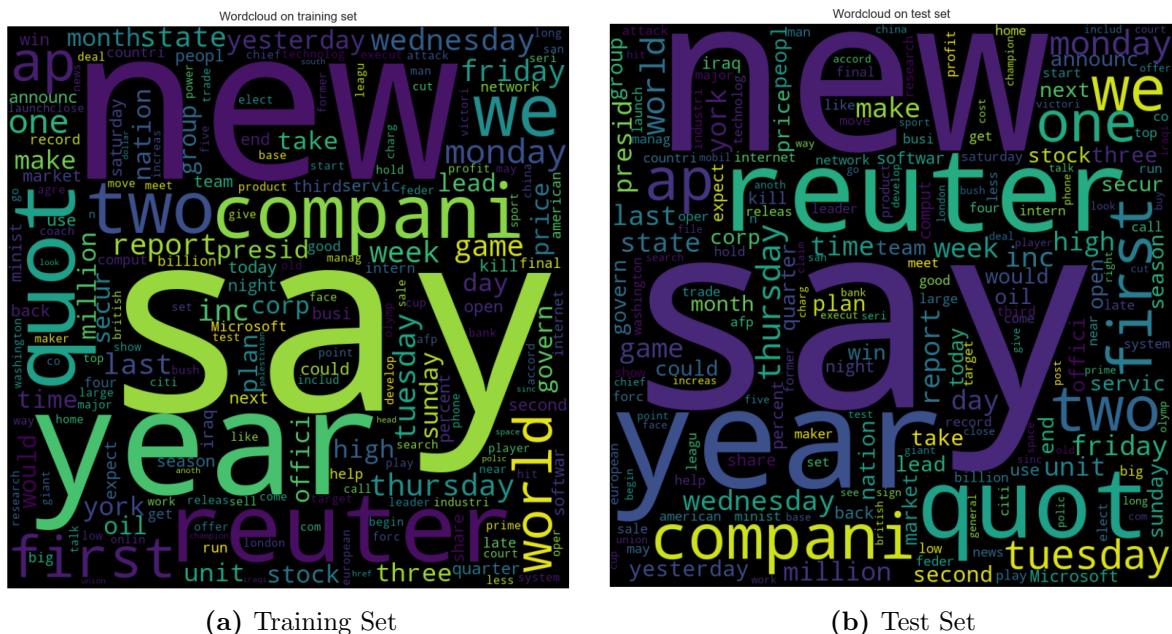


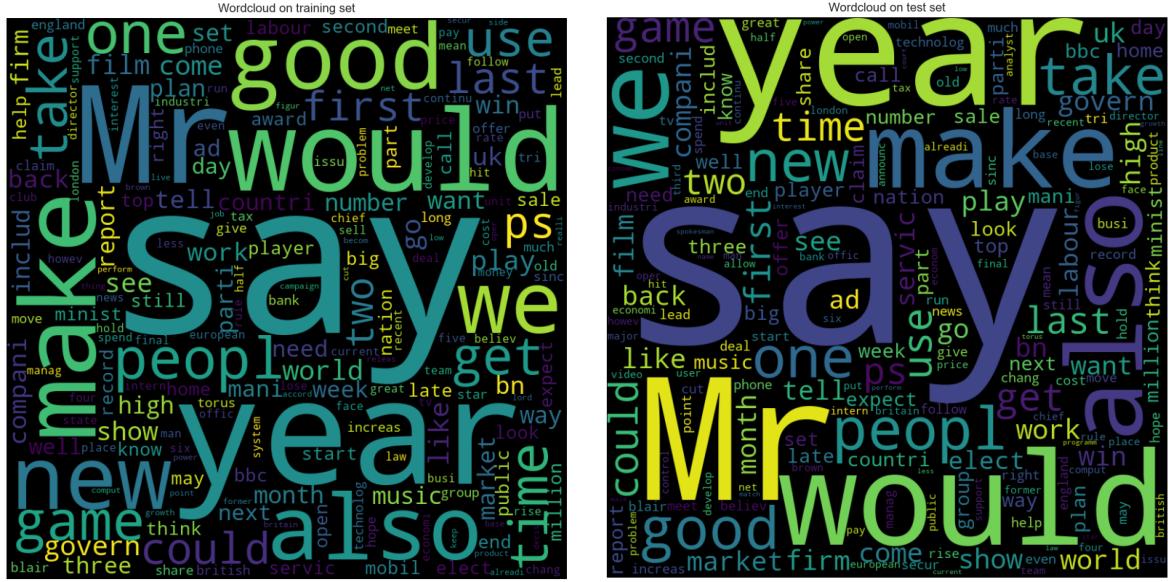
Figure 3: Wordcloud AG News

⁶Link for Texthero library: <https://texthero.org>

⁷Link for Nltk library: <https://www.nltk.org>

⁸Link for Contractions library: <https://libraries.io/pypi/contractions/0.1.73>

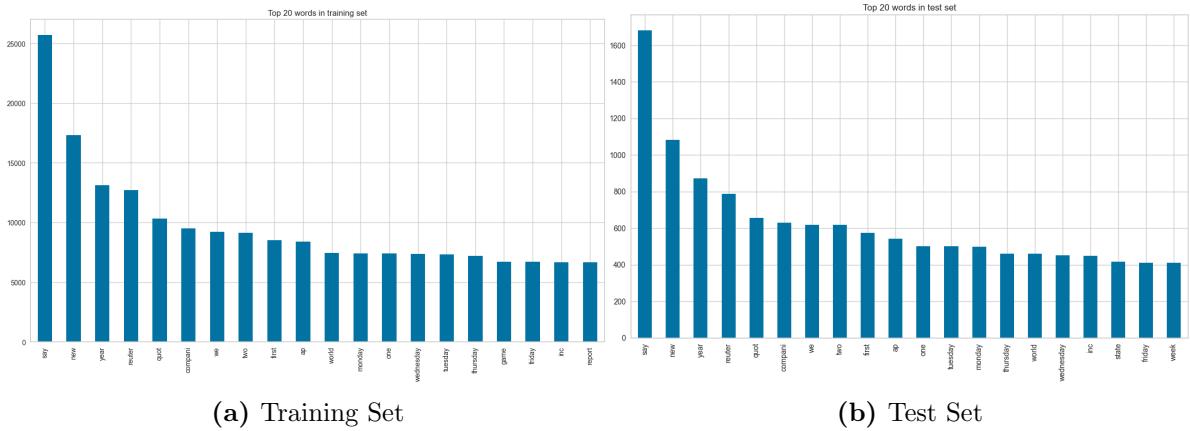
⁹Link for Simplemma library: <https://libraries.io/pypi/simplemma>



(a) Training Set

(b) Test Set

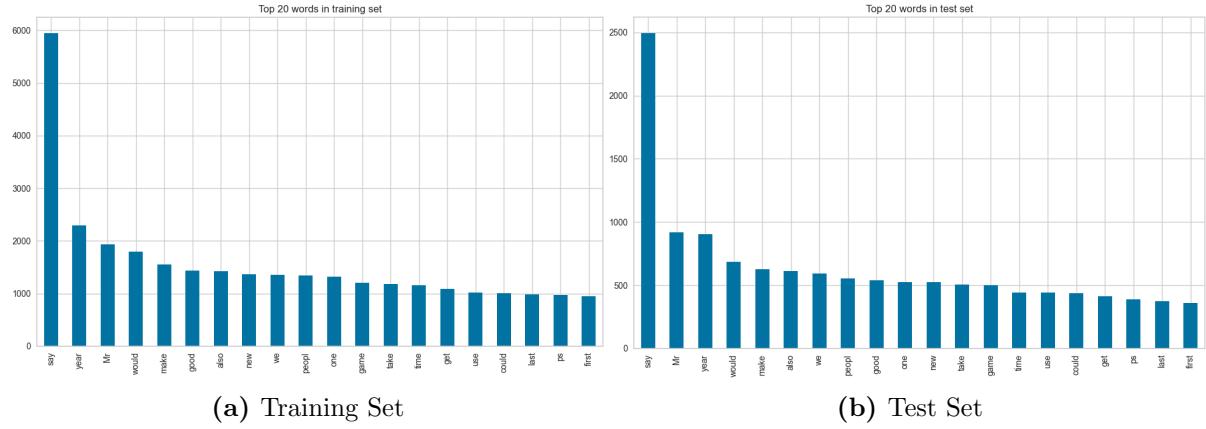
Figure 4: Wordcloud BBC News



(a) Training Set

(b) Test Set

Figure 5: Top 20 Words on AG News



(a) Training Set

(b) Test Set

Figure 6: Top 20 Words on BBC News

```

Information on Dataset

Before Preprocess:
The first rows of training set are:
0 Reuters - Short-sellers, Wall Street's dwindle...
1 Reuters - Private investment firm Carlyle Grou...
2 Reuters - Soaring crude prices plus worries ab...
3 Reuters - Authorities have halted oil export...
4 AFP - Tearaway world oil prices, topping reco...
Name: Description, dtype: object

The first rows of training set are:
0 3
1 3
2 3
3 3
4 3
Name: Class Index, dtype: int64

The first rows of test set are:
0 Unions representing workers at Turner Newall...
1 SPACER TORONTO, Canada -- A second team o...
2 AP - A company founded by a chemistry research...
3 AP - It's barely dawn when Mike Fitzpatrick st...
4 AP - Southern California's smog-fighting agenc...
Name: Description, dtype: object

The first rows of test set are:
0 3
1 4
2 4
3 4
4 4
Name: Class Index, dtype: int64

The class distribution on training set is
Business 30000
Sci/Tech 30000
Sports 30000
World 30000
Name: Class Index, dtype: int64

The class distribution on test set is
Business 1900
Sci/Tech 1900
Sports 1900
World 1900
Name: Class Index, dtype: int64

The numbers of Nan value on training set are 0
The numbers of Nan value on test set are 0
The numbers of duplicate elements on training set are 1179
The numbers of duplicate elements on test set are 6
There are 120000 rows in the training set
There are 7600 rows in the test set

After Preprocess:
The first rows of training set are:
0 reuter short seller wall street dwindle band ul...
1 reuter privat invest firm carlyle group reput...
2 reuter soar crude price plus worri economi out...
3 reuter author halt oil export flow main pipeli...
4 afp tearaway world oil price toppi record stra...
Name: Description, dtype: object

The first rows of training set are:
0 3
1 3
2 3
3 3
4 3
Name: Class Index, dtype: int64

The first rows of test set are:
0 3
1 4
2 4
3 4
4 4
Name: Class Index, dtype: int64

The class distribution on training set is
Sports 29837
World 29778
Business 29646
Sci/Tech 29560
Name: Class Index, dtype: int64

The class distribution on test set is
World 1900
Sci/Tech 1899
Sports 1899
Business 1896
Name: Class Index, dtype: int64

The numbers of Nan value on training set are 0
The numbers of Nan value on test set are 0
The numbers of duplicate elements on training set are 0
The numbers of duplicate elements on test set are 0
There are 118821 rows in the training set
There are 7594 rows in the test set

```

(a) AG News

```

Information on Dataset

Before Preprocess:
The first rows of training set are:
0 tv future in the hands of viewers with home th...
1 worldcom boss left books alone former worldc...
2 tigers wary of farrell gamble leicester say ...
3 yeading face newcastle in fa cup premiership...
4 ocean s twelve raids box office ocean s twelve...
Name: text, dtype: object

The first rows of training set are:
0 tech
1 business
2 sport
3 sport
4 entertainment
Name: category, dtype: object

There is no test set to calculate first rows

The class distribution on training set is
sport 511
business 510
politics 417
tech 401
entertainment 386
Name: category, dtype: int64

There is no test set to calculate class distribution
The numbers of Nan value on training set are 0
There is no test set to calculate Nan value
There are 2225 rows in the training set
There is no test set to calculate number of rows
The numbers of duplicate elements on training set are 99
There is no test set to calculate number of duplicates
There are 1488 rows in the training set
There is no test set to calculate number of duplicates
There are 638 rows in the test set

After Preprocess:
The first rows of training set are:
1014 howard dismiss torus tax fear michael howard d...
1675 johnni cash manag hollli die former manag john...
178 yuko owner sue russia 28bn major owner embattl...
527 mobil doubl bus ticket mobil could soon doubl...
753 itali aim ratti england itali coach john kirwa...
Name: text, dtype: object

The first rows of training set are:
0 2
1 1
2 0
3 4
4 3
dtype: int64

The first rows of test set are:
664 uk premier ring music produc behind lord ring ...
1801 continent may run cash share continent airlin ...
1258 honda win china copyright rule japan honda cop...
1881 woolf murder sentenc rethink plan give murde...
839 format war could contus user technolog firm so...
Name: text, dtype: object

The first rows of test set are:
0 1
1 0
2 0
3 2
4 4
dtype: int64

The class distribution on training set is
sport 504
business 503
politics 403
entertainment 369
tech 347
Name: category, dtype: int64

There is no test set to calculate class distribution
The numbers of Nan value on training set are 0
There is no test set to calculate Nan value
There are 1488 rows in the training set
There is no test set to calculate number of duplicates
There are 638 rows in the test set

```

(b) BBC News

Figure 7: Before and After Preprocess

3 — Classification

After the preprocessing phase, the dataset is ready to be used to learn classification models that will be used in the final application. In this chapter are discussed the chosen strategies relative to classification phase. Before apply classification is necessary to transform text into vector of real numbers, which is the format that ML models support. The process to convert text data into numerical data/vector, is called vectorization.

3.1 Vectorization

The solution used to implement vectorization is **Term Frequency-Inverse Document Frequencies (TF-IDF)**¹⁰. It is a numerical statistic that's intended to reflect how important a word is to a document. Words that get repeated too often don't overpower less frequent but important words. It is composed by two parts:

1. **Term Frequency (TF)**. It can be understood as a normalized frequency score and it is always ≤ 1 . It is calculated via the following formula:

$$TF = \frac{\text{Frequency of word in a document}}{\text{Total number of words in that document}}$$

2. **Inverse Document Frequency (IDF)**, but before is necessary make sense of $DF - \text{Document Frequency}$. It's given by the following formula:

$$DF(\text{word}) = \frac{\text{Number of documents with word in it}}{\text{Total number of documents}}$$

It measures the proportion of documents that contain a certain word. IDF is the reciprocal of the Document Frequency, and the final IDF score comes out of the following formula:

$$IDF(\text{word}) = \log \left(\frac{\text{Total number of documents}}{\text{Number of documents with word in it}} \right)$$

The intuition behind it is that the more common a word is across all documents, the lesser its importance is for the current document. A logarithm is taken to dampen the effect of IDF in the final calculation. The final $TF-IDF$ score comes out to be:

$$TF - IDF = TF \cdot IDF$$

The higher the score and more important that word is. Basically, the value of a word increases proportionally to count in the document, but it is inversely proportional to the frequency of the word in the corpus.

3.2 Classifiers

It's time to train a machine learning models on the vectorized dataset. To minimize lengthy re-training and allow you to share, commit, and re-load pre-trained machine learning models is used *Dill library*¹¹, that is a useful Python tool that allows to save ML models. Every times that application starts, it controls if models are present in *models_saved* directory. If a model is present, it is being loaded otherwise it is being trained and saved in *models_saved* folder.

¹⁰Link for TfidfVectorizer library: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

¹¹Link for Dill library: <https://libraries.io/pypi/dill>

Another important things is that, every times that the user change a setting (for example changing dataset), all models are primarily removed, then trained and at the end saved.

To find the optimal parameters from the chosen classifiers is performed a technique called *GridSearchCV*¹². The performance of a model significantly depends on the value of hyperparameters. There is no way to know in advance the best values for hyperparameters so ideally, is necessary to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus a solution is use GridSearchCV to automate the tuning of hyperparameters.

In this application the chosen classifiers are:

- *MultinomialNB*¹³
- *Logistic Regression*¹⁴
- *SGD Classifier* ¹⁵

3.2.1 Hyperparameters

This section is focused on GridSearchCV and hyperparameters of its classifiers. The parameters of the estimator used are optimized by cross-validated using *StratifiedKFold*¹⁶ over a parameter grid. The parameters grid for each classifier is:

- *MultinomialNB*
 - alpha: (1, 0.8, 0.7, 0.5, 0.3, 0.1, 0.05, 0.01, 0.001, 0.0001, 0.00001)
- *Logistic Regression*
 - C : [25, 20, 15, 10, 5, 3, 1, 0.1, 0.05, 0.01]
 - solver: ['liblinear', 'newton-cg']
- *SGD Classifier*
 - eta0: [0.0, 0.03, 0.01, 0.003, 0.001, 0.0003],
 - penalty: ['l1', 'l2', 'elasticnet']
 - alpha: [1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001, 0.0003, 0.0001]
 - loss: ['log_loss', 'modified_huber']

Here are reported the best parameters, best score and time to the fit classifiers for each dataset:

¹²Link for GridSearchCV library: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

¹³Link for MultinomialNB library: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

¹⁴Link for Logistic Regression library: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

¹⁵Link for SGDClassifier library: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

¹⁶Link for StratifiedKFold library: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

AG News

- *MultinomialNB*
 - best parameters
 - * alpha: 0.3
 - best score: 0.894
 - time: 35.087 seconds
- *Logistic Regression*
 - best parameters
 - * C: 3
 - * solver: 'liblinear'
 - best score: 0.9
 - time: 402.227 seconds
- *SGD Classifier*
 - best parameters
 - * eta0: 0.0
 - * penalty: 'l2'
 - * alpha: 0.0001
 - * loss: 'modified_huber'
 - best score: 0.901
 - time: 1121.933 seconds

BBC News

- *MultinomialNB*
 - best parameters
 - * alpha: 0.8
 - best score: 0.976
 - time: 5.245 seconds
- *Logistic Regression*
 - best parameters
 - * C: 25
 - * solver: 'liblinear'
 - best score: 0.98
 - time: 14.081 seconds
- *SGD Classifier*
 - best parameters
 - * eta0: 0.0
 - * penalty: 'l2'
 - * alpha: 0.003
 - * loss: 'modified_huber'
 - best score: 0.98
 - time: 63.516 seconds

The time to fit all classifiers is calculated using *Apple M1* processor.

4 — Results

In this chapter there is the description of the results obtained using following statistics:

- *Classification Report*¹⁷
- *Confusion Matrix*¹⁸
- *Area under the Curve*¹⁹
- *Class Prediction Error*²⁰

The following sections report results obtained for each dataset.

4.1 AG News

4.1.1 Classification Report

MultinomialNB				
	Precision	Recall	F1-Score	Support
World	0.92	0.89	0.90	1900
Sports	0.94	0.97	0.96	1899
Business	0.87	0.84	0.85	1896
Sci/Tech	0.85	0.88	0.87	1899
Accuracy			0.90	7594
Macro Avg	0.89	0.90	0.89	7594
Weighted Avg	0.90	0.90	0.89	7594

Final Training Accuracy: 91% Model Accuracy: 90%

Logistic Regression				
	Precision	Recall	F1-Score	Support
World	0.93	0.90	0.92	1900
Sports	0.95	0.98	0.96	1899
Business	0.88	0.88	0.88	1896
Sci/Tech	0.89	0.89	0.89	1899
Accuracy			0.91	7594
Macro Avg	0.91	0.91	0.91	7594
Weighted Avg	0.91	0.91	0.91	7594

Final Training Accuracy: 95% Model Accuracy: 91%

¹⁷Link for Classification Report library: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

¹⁸Link for ConfusionMatrixDisplay library: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ConfusionMatrixDisplay.html>

¹⁹Link for ROCAUC library: <https://www.scikit-yb.org/en/latest/api/classifier/rocauc.html>

²⁰Link for ClassPredictionError library: https://www.scikit-yb.org/en/latest/api/classifier/class_prediction_error.html

SGD Classifier					
	Precision	Recall	F1-Score	Support	
World	0.93	0.90	0.91	1900	
Sports	0.94	0.98	0.96	1899	
Business	0.88	0.87	0.88	1896	
Sci/Tech	0.88	0.88	0.88	1899	
Accuracy			0.91	7594	
Macro Avg	0.91	0.91	0.91	7594	
Weighted Avg	0.91	0.91	0.91	7594	

Final Training Accuracy: 93% Model Accuracy: 91%

4.1.2 Confusion Matrix

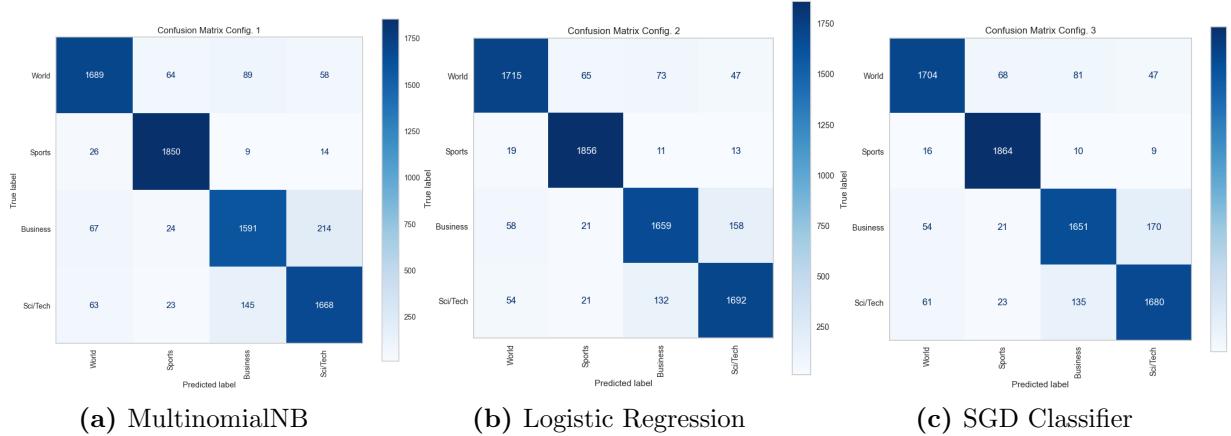


Figure 8: Confusion Matrix AG News

4.1.3 Area Under the Curve

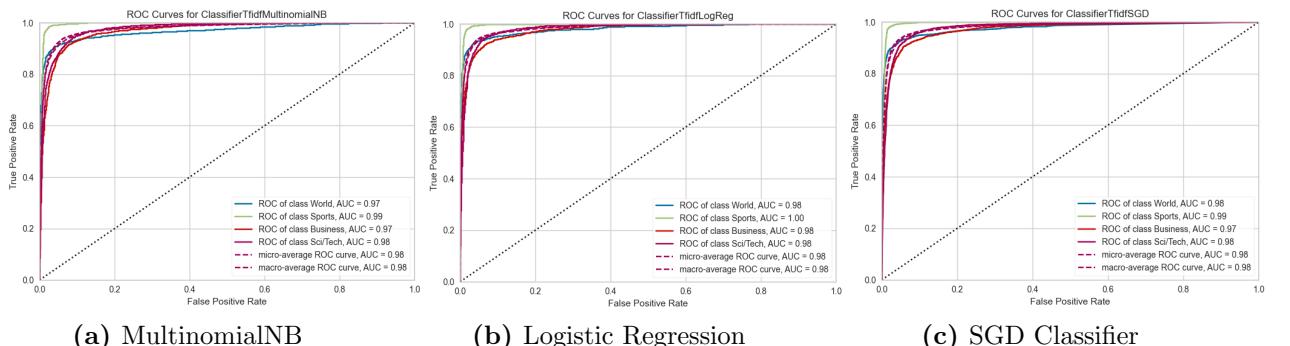


Figure 9: Area Under the Curve AG News

4.1.4 Class Prediction Error

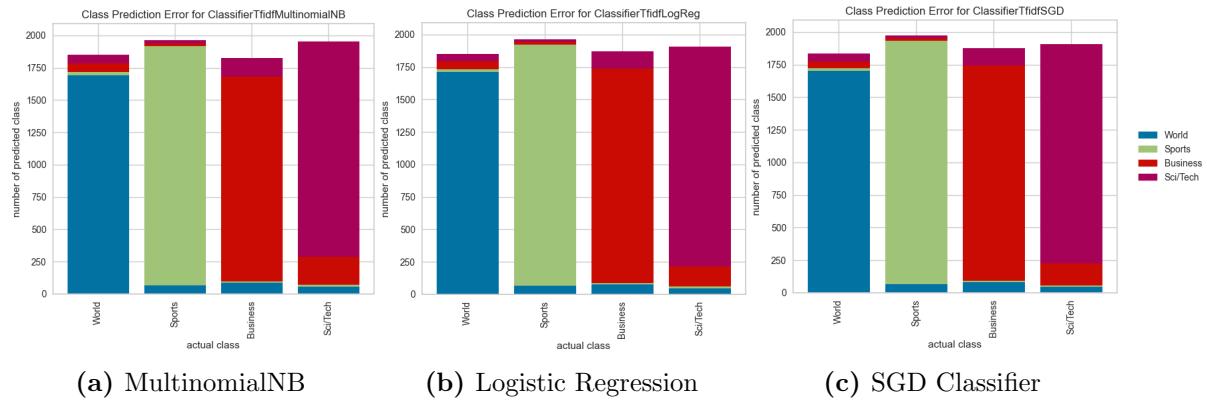


Figure 10: Class Prediction Error AG News

4.2 BBC News

4.2.1 Classification Report

MultinomialNB				
	Precision	Recall	F1-Score	Support
business	0.96	0.98	0.97	162
entertainment	0.99	0.92	0.95	96
politics	0.95	0.98	0.97	128
sport	0.99	1.00	1.00	141
tech	0.95	0.95	0.95	111
Accuracy			0.97	638
Macro Avg	0.97	0.96	0.97	638
Weighted Avg	0.97	0.97	0.97	638

Final Training Accuracy: 99% Model Accuracy: 97%

Logistic Regression				
	Precision	Recall	F1-Score	Support
business	0.97	0.98	0.98	162
entertainment	0.99	0.99	0.99	96
politics	0.96	0.98	0.97	128
sport	0.99	1.00	1.00	141
tech	1.00	0.95	0.98	111
Accuracy			0.98	638
Macro Avg	0.98	0.98	0.98	638
Weighted Avg	0.98	0.98	0.98	638

Final Training Accuracy: 100% Model Accuracy: 98%

SGD Classifier					
	Precision	Recall	F1-Score	Support	
business	0.96	0.98	0.97	162	
entertainment	0.99	0.99	0.99	96	
politics	0.96	0.98	0.97	128	
sport	0.99	1.00	0.99	141	
tech	1.00	0.94	0.97	111	
Accuracy			0.98	638	
Macro Avg	0.99	0.98	0.98	638	
Weighted Avg	0.98	0.98	0.98	638	

Final Training Accuracy: 100% Model Accuracy: 98%

4.2.2 Confusion Matrix

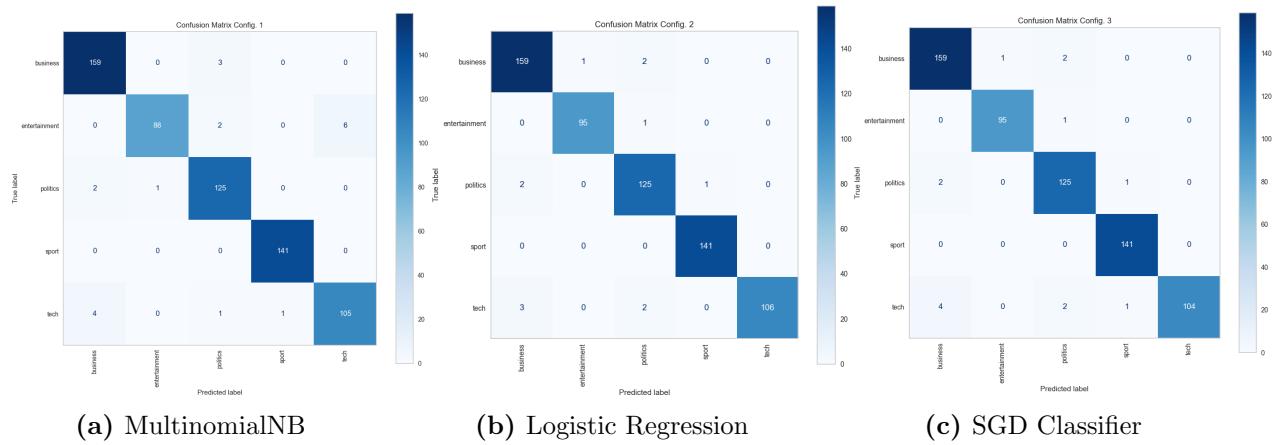


Figure 11: Confusion Matrix BBC News

4.2.3 Area Under the Curve

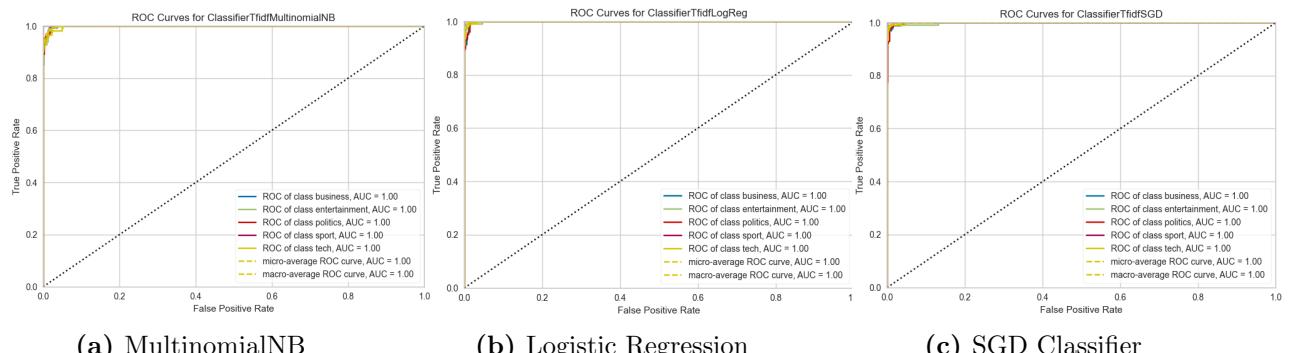


Figure 12: Area Under the Curve BBC News

4.2.4 Class Prediction Error

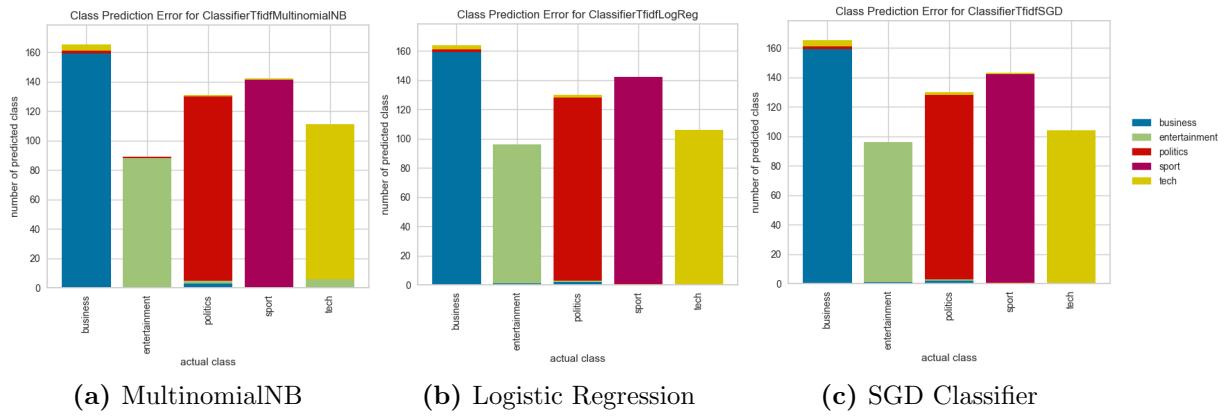


Figure 13: Class Prediction Error BBC News

5 — Application

News Classifier is a simple application based on text mining in order to classify news articles to their right topic.

5.1 How to Install

Before use the application is necessary to install all library used using this command: `pip install -r requirements.txt`. After installing all libraries, the command to start application is `python ./src/main.py`. The application is tested with `python 3.8`.

5.2 How to classify text

Every user can operate with the applicative as they open it and insert into the textbox a news to classify.

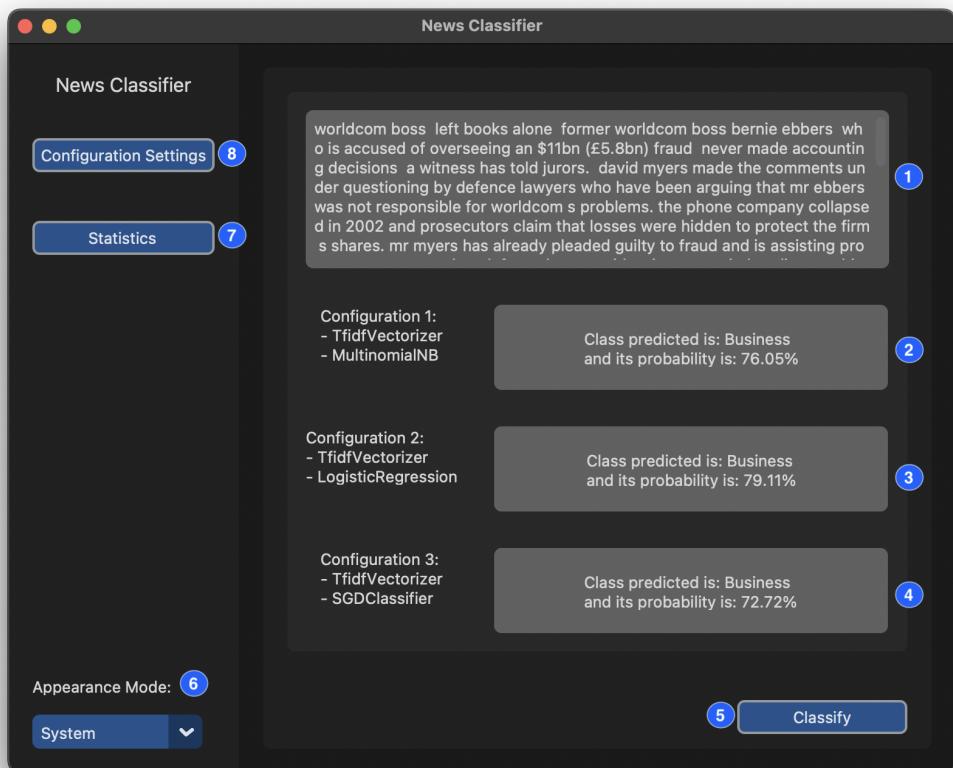


Figure 14: Home Application

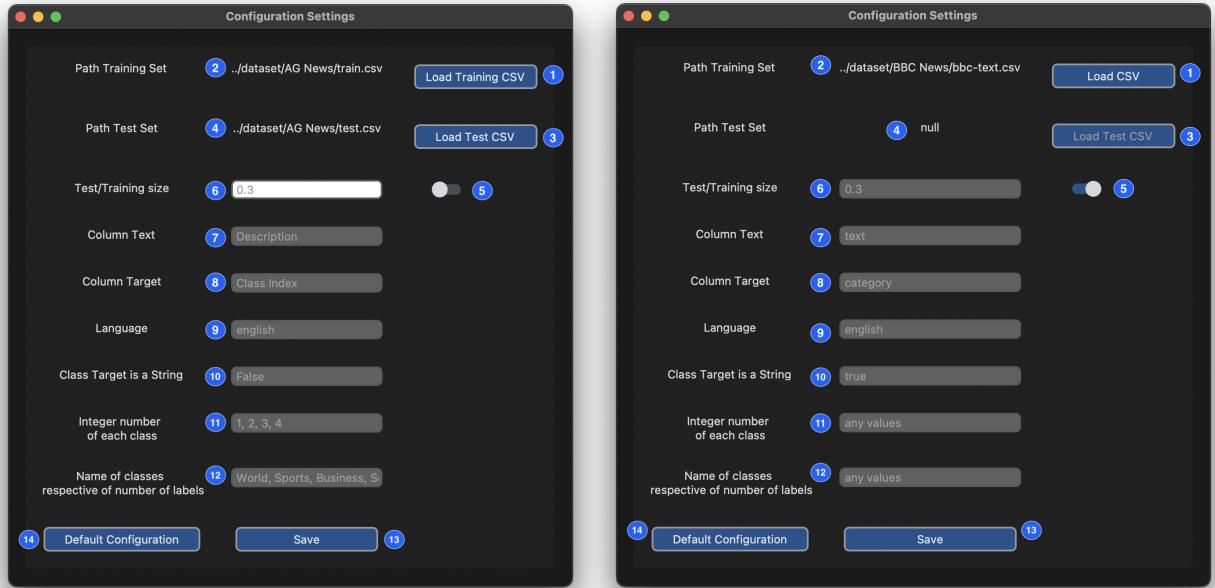
1. *TextBox*: Input text to classify
2. *Prediction first classifier*: Output text displays class predicted, by the first classifier, and its probability
3. *Prediction second classifier*: Output text displays class predicted, by the second classifier, and its probability

4. *Prediction third classifier*: Output text displays class predicted by the third classifier, and its probability
5. *Classify*: Button to classify text inserted in *TextBox*
6. *Appearance*: Option menu to change the appearance of the application (Dark, Light and System)
7. *Statistics*: Button to see all statistics on dataset and classifiers
8. *Configuration Settings*: Button to change dataset

5.3 How to change dataset

If the user wants to change dataset, the steps are:

- Select one or two *.csv files* as input
- Insert the configuration values for relative dataset given in input



(a) Configuration AG News

(b) Configuration BBC News

Figure 15: Configuration

1. *Load Training CSV/Load CSV*: Button to import training or dataset file
2. *Path Training Set*: Text to display the path of training file
3. *Load Test CSV*: Button to import test file (if switch is on)
4. *Path Test Set*: Text to display the path of test file or null if the dataset is composed by one file
5. *Switch*: Status on to load dataset composed by only one file and set *Test/Training Size* value. Status off load trainig and test files

6. *Test/Training Size*: Value of the test size (only if there is only one file)
7. *Column Text*: CSV column name to detect text to classify
8. *Column Target*: CSV column name contains classes
9. *Language*: Text to set language of the text to classify
10. *Class Target is a String*: Boolean value: true if classes are described by strings, false by number
11. *Integer Number of Each Class*: Text to set the list of numeric classes when *Class Target is a String* is false
12. *Name of classes respective of number of label*: Text to set the list of classes name when *Class Target is a String* is false
13. *Save*: Button to save settings
14. *Default Configuration*: Button to restore default settings

5.4 Display statistics

The image below describes all statistics in the application, each buttons display the corresponding statistics.

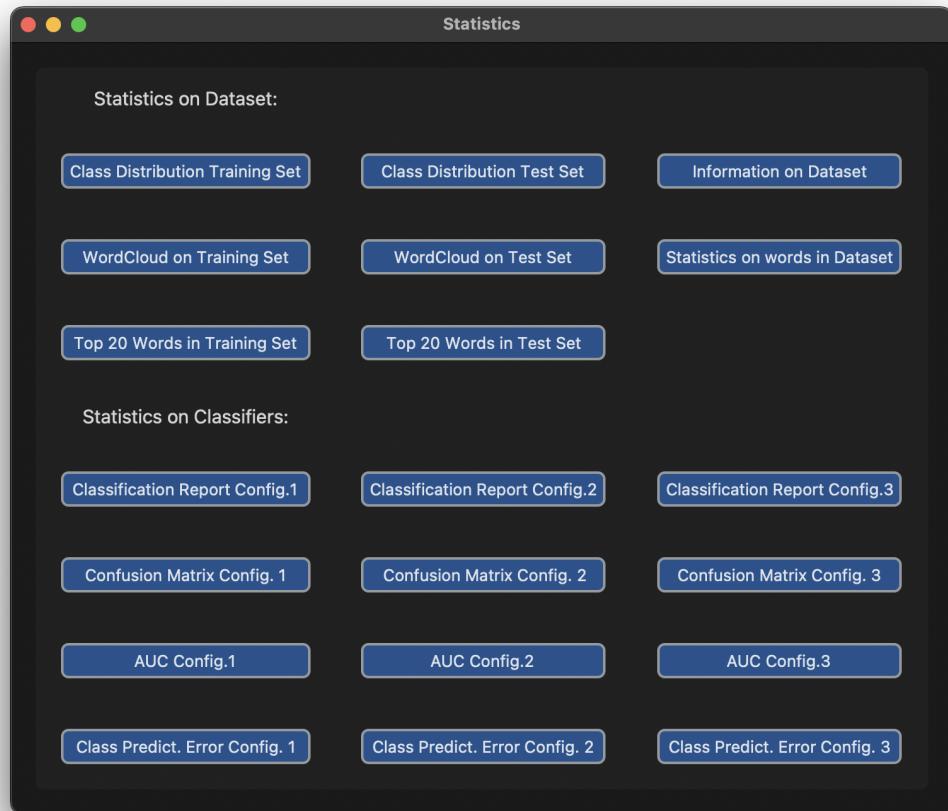


Figure 16: Statistics