

LEONARDO BARGIOTTI

NEWS CLASSIFIER

DATA MINING AND MACHINE LEARNING

PROJECT GOALS

News Classifier is an application that exploit text mining to categorize news with the right topic.

The application can work with different dataset, but they must have the same structure, that is: one column for text in input and another one for the true class.



DATASET DESCRIPTION

AG News

- **Source:**
https://www.kaggle.com/datasets/amana_nandrai/ag-news-classification-dataset?select=train.csv
- **Classes:** "world", "sports", "business", and "sci/tech"
- **Volume:** training samples is 120,000 and testing 7,600.

BBC News

- **Source:**
<https://www.kaggle.com/code/rockystats/bbc-text-classification-word2vec-vs-tf-idf/data>
- **Classes:** "tech", "business", "sport", "entertainment" and "politics"
- **Volume:** dataset samples is 2225.

NEWS CLASSIFICATION



PRE-PROCESSING



LEARNING PHASE



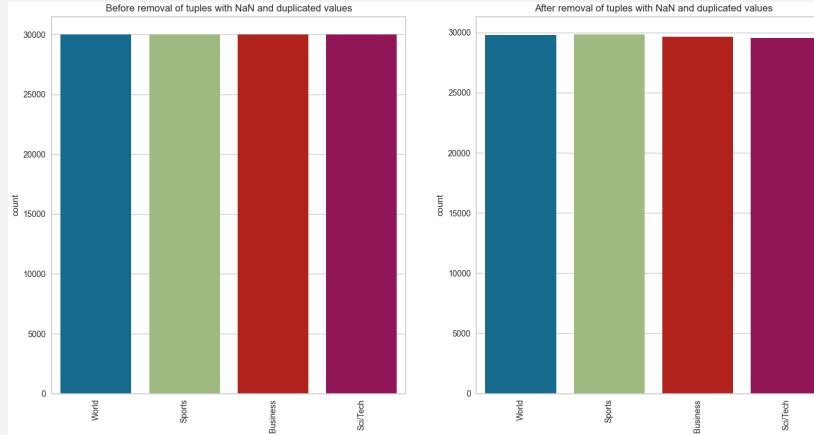
RESULTS



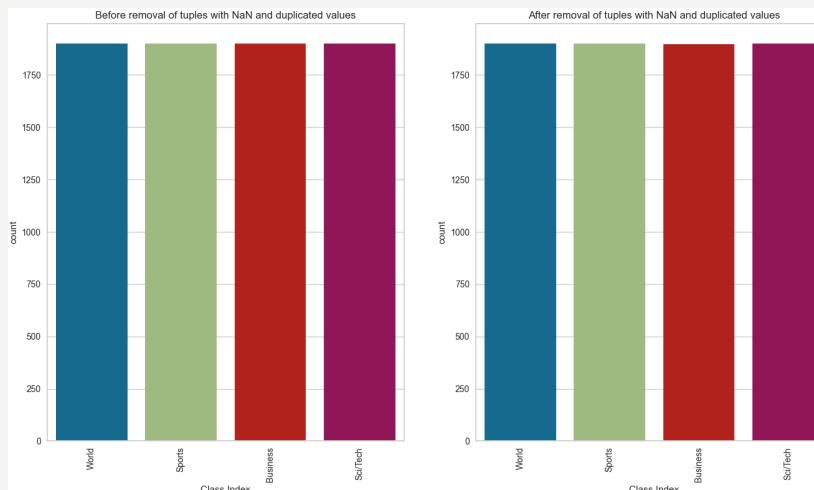
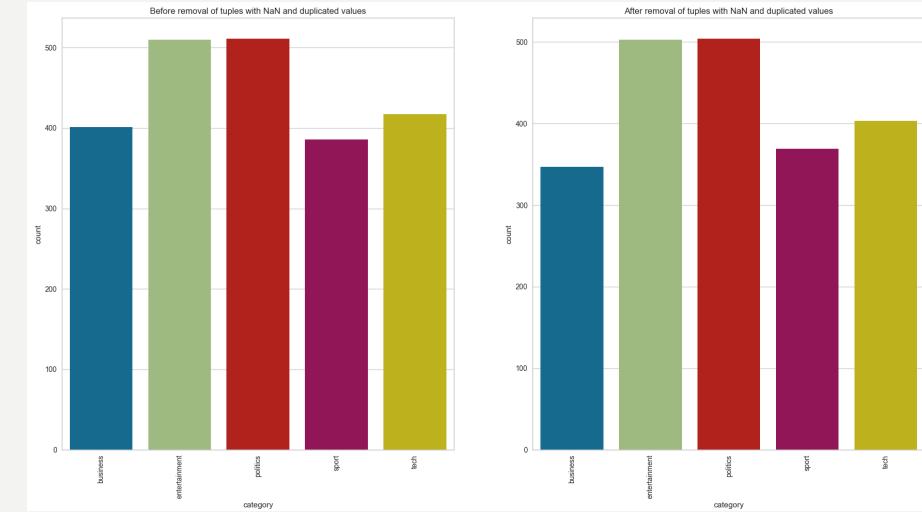
PRE- PROCESSING

MOVING DUPLICATES AND MISSING VALUE

AG News



BBC NEWS



TEXT PRE-PROCESS

PRE-PROCESS

- Apply:
 - Lowercase
 - Expanding Contractions
- Removing:
 - Stopwords
 - Digits
 - Punctuation
 - Diactrics
 - HTML Tags
 - URLs
 - Extra Whitespace

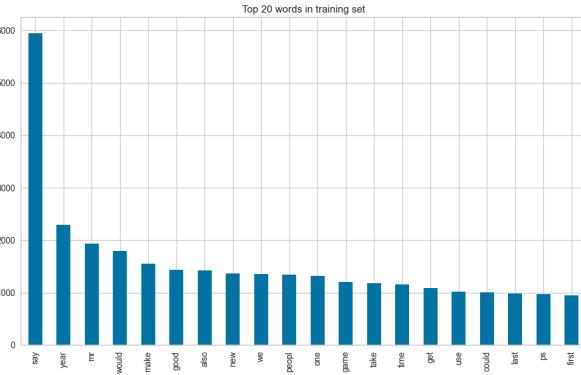
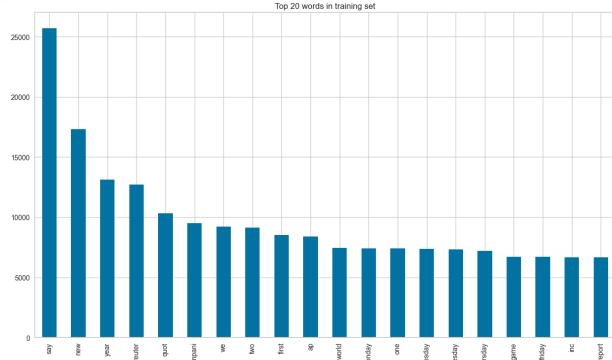
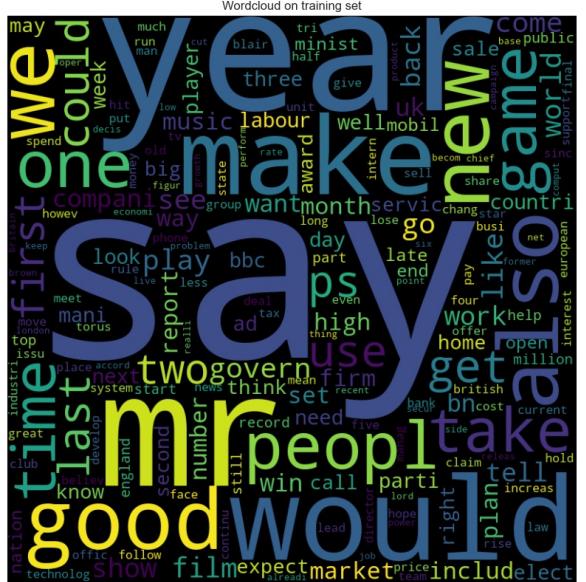
STEMMING

Stemming is the process of reducing words to their root form.

LEMMATIZATION

The goal of lemmatization is the same as for stemming.

Lemmatization, on the other hand, is a tool that performs full morphological analysis to more accurately find the root.





LEARNING PHASE

CLASSIFICATION

The goal of **TfidfVectorizer** is to convert text data into numerical data that's intended to reflect how important a word is to a document.

Tf is the number of times a term appears in a particular document, **Idf** is a measure of how common or rare a term is across the entire corpus of documents.

The higher is the value, the more relevant the term is in that document.

- Classifiers:

- MultinomialNB
 - alpha: [1, 0.9, 0.8, 0.7, 0.6, 0.5, , 0.4, 0.3, 0.1, 0.05, 0.01, 0.001, 0.0001, 0.00001]
- Logistic Regression
 - C : [100, 75, 50, 2515, 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']



AG News

- MultinomialNB
 - alpha: 0.3
 - best score: 0.895
- Logistic Regression
 - C: 3, solver: 'liblinear'
 - best score: 0.9
- SGD Classifier
 - eta0: 0.0, penalty: 'l2', alpha: 0.0001, loss: 'modified huber'
 - best score: 0.901

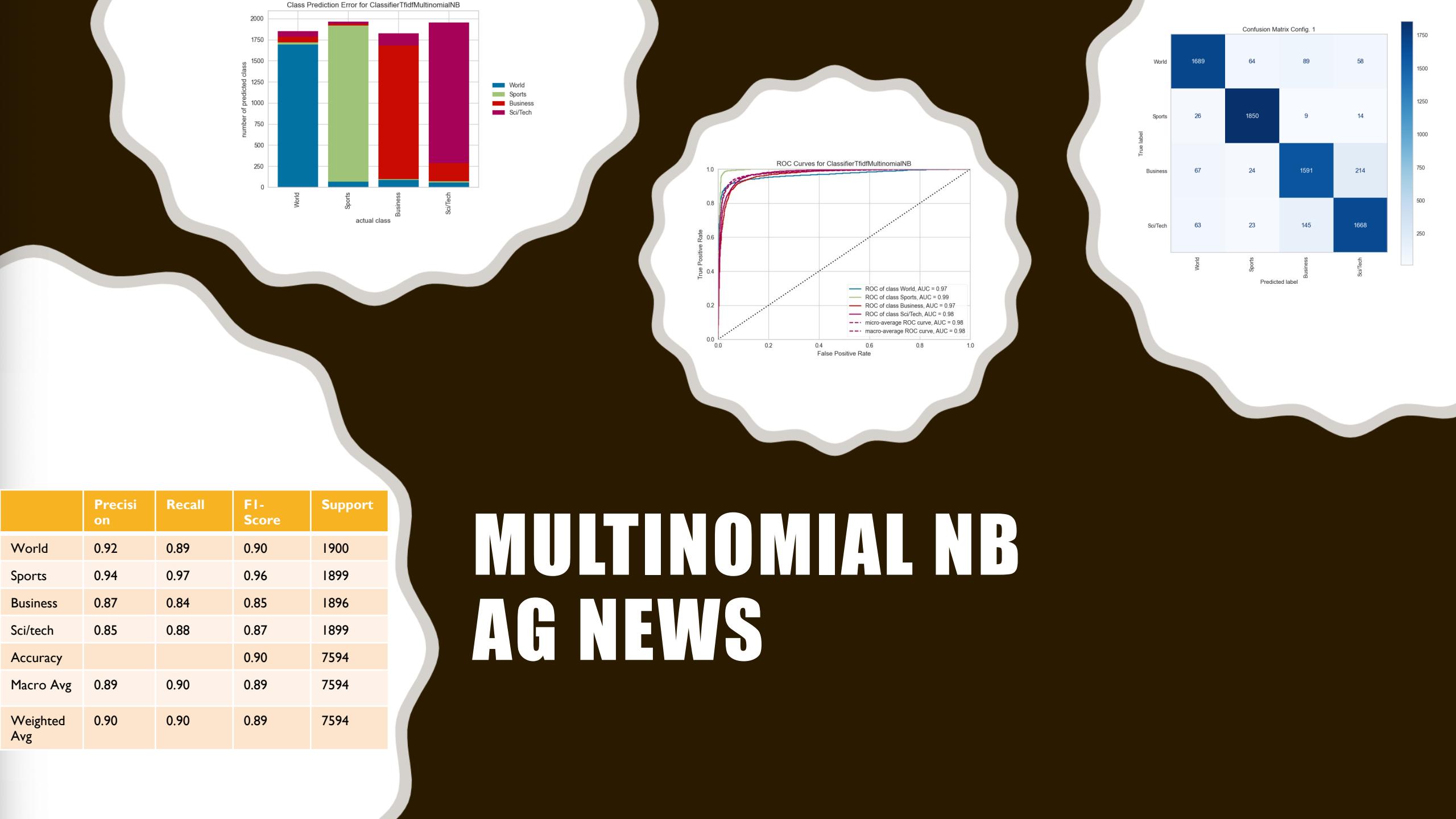


BBC News

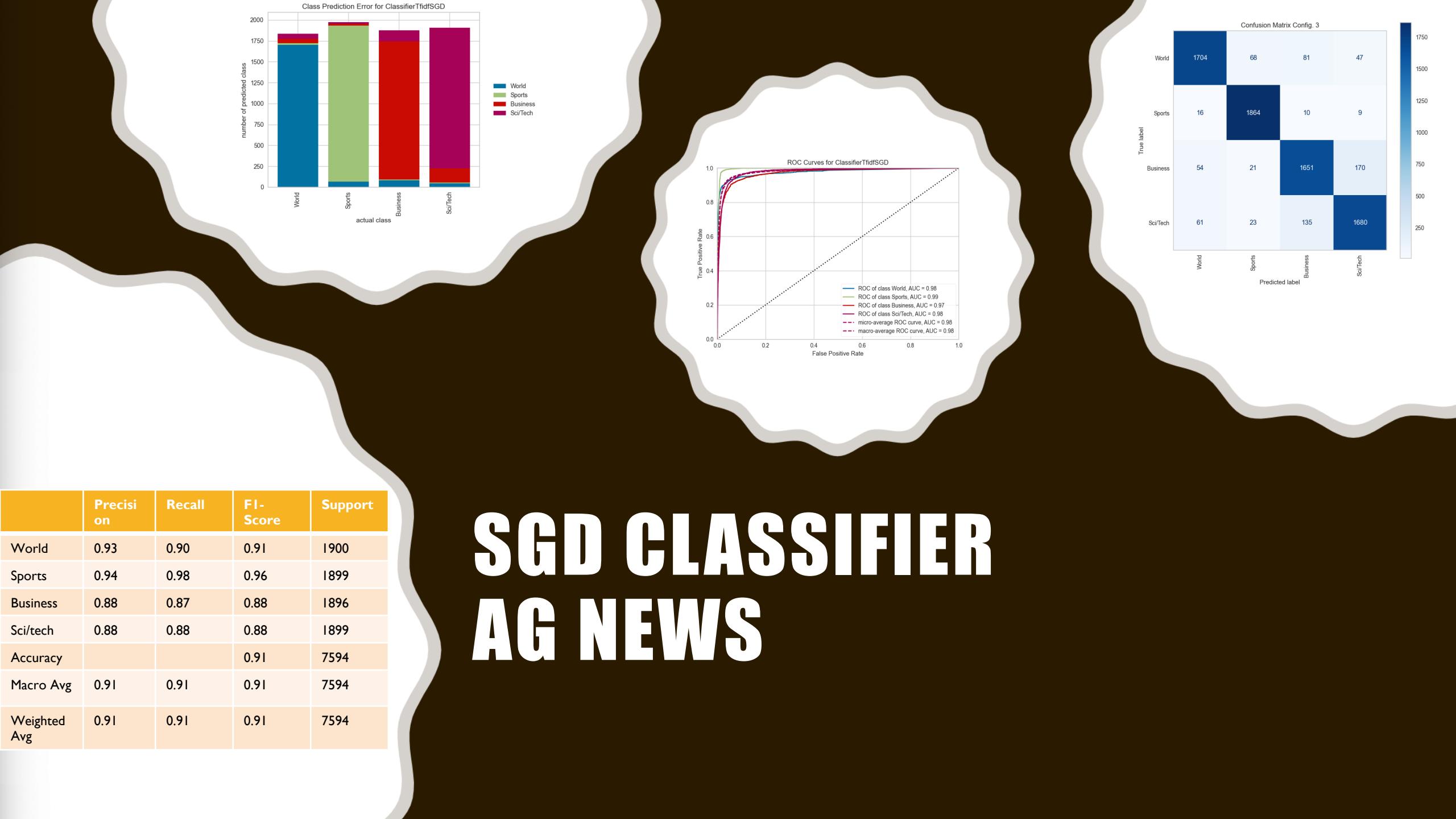
- MultinomialNB
 - alpha: 0.7
 - best score: 0.976
- Logistic Regression
 - C: 100, solver: 'liblinear'
 - best score: 0.981
- SGD Classifier
 - eta0: 0.0, penalty: 'l2', alpha: 0.003, loss: 'modified huber'
 - best score: 0.98

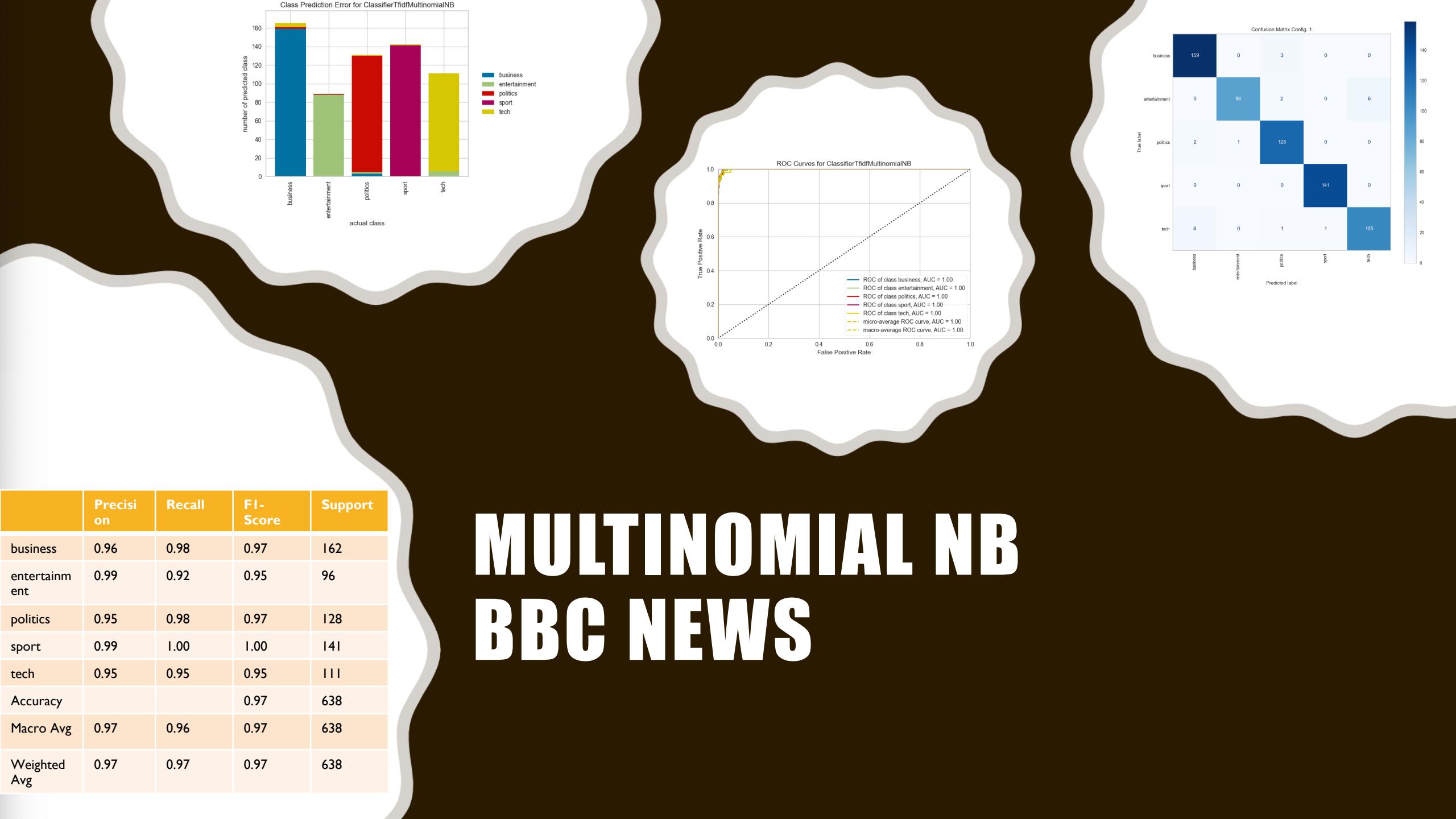


RESULTS

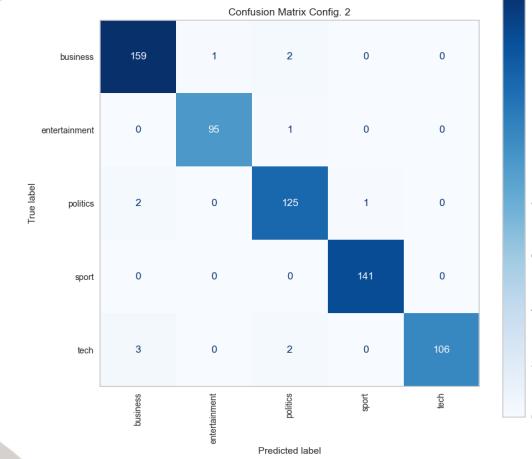
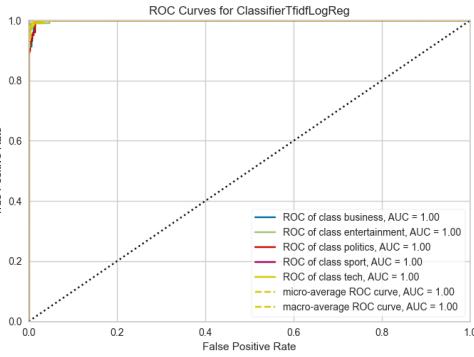
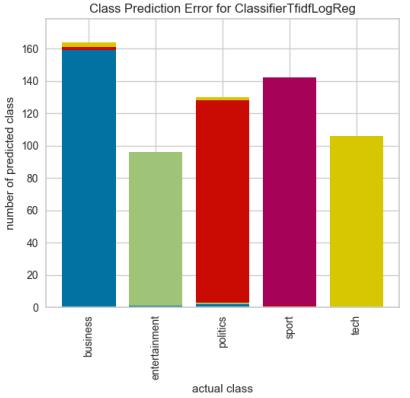






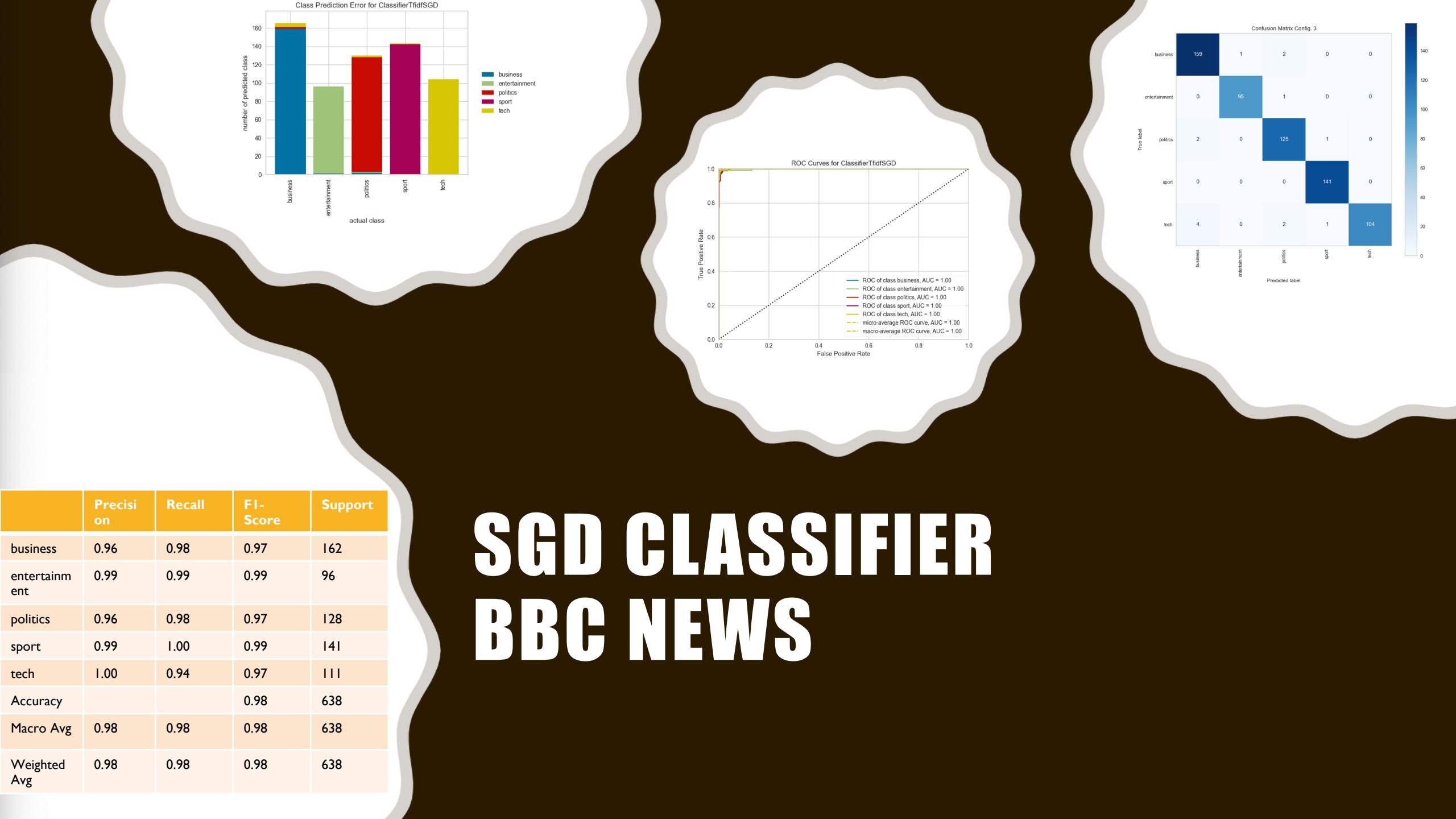


	Precision	Recall	F1-Score	Support
business	0.96	0.98	0.98	162
entertainment	0.99	0.99	0.99	96
politics	0.96	0.98	0.97	128
sport	0.99	1.00	1.00	141
tech	1.00	0.95	0.95	111
Accuracy			0.98	638
Macro Avg	0.98	0.98	0.98	638
Weighted Avg	0.98	0.98	0.98	638



LOGISTIC REGRESSION

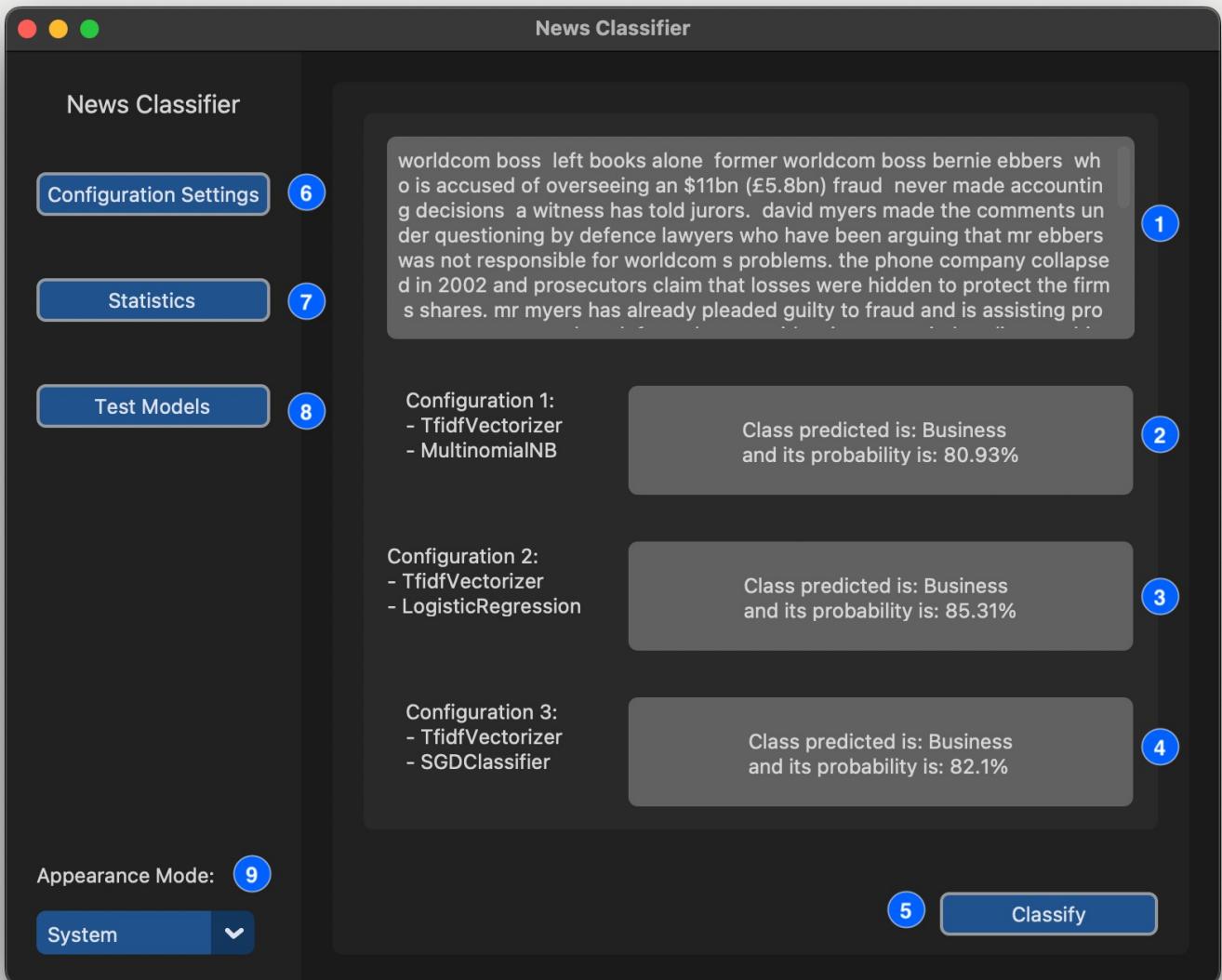
BBC NEWS





APPLICATION GUI

APPLICATION HOME



1 Text to classify

2 Class predicted by first classifier

3 Class predicted by second classifier

4 Class predicted by third classifier

5 Button to classify text

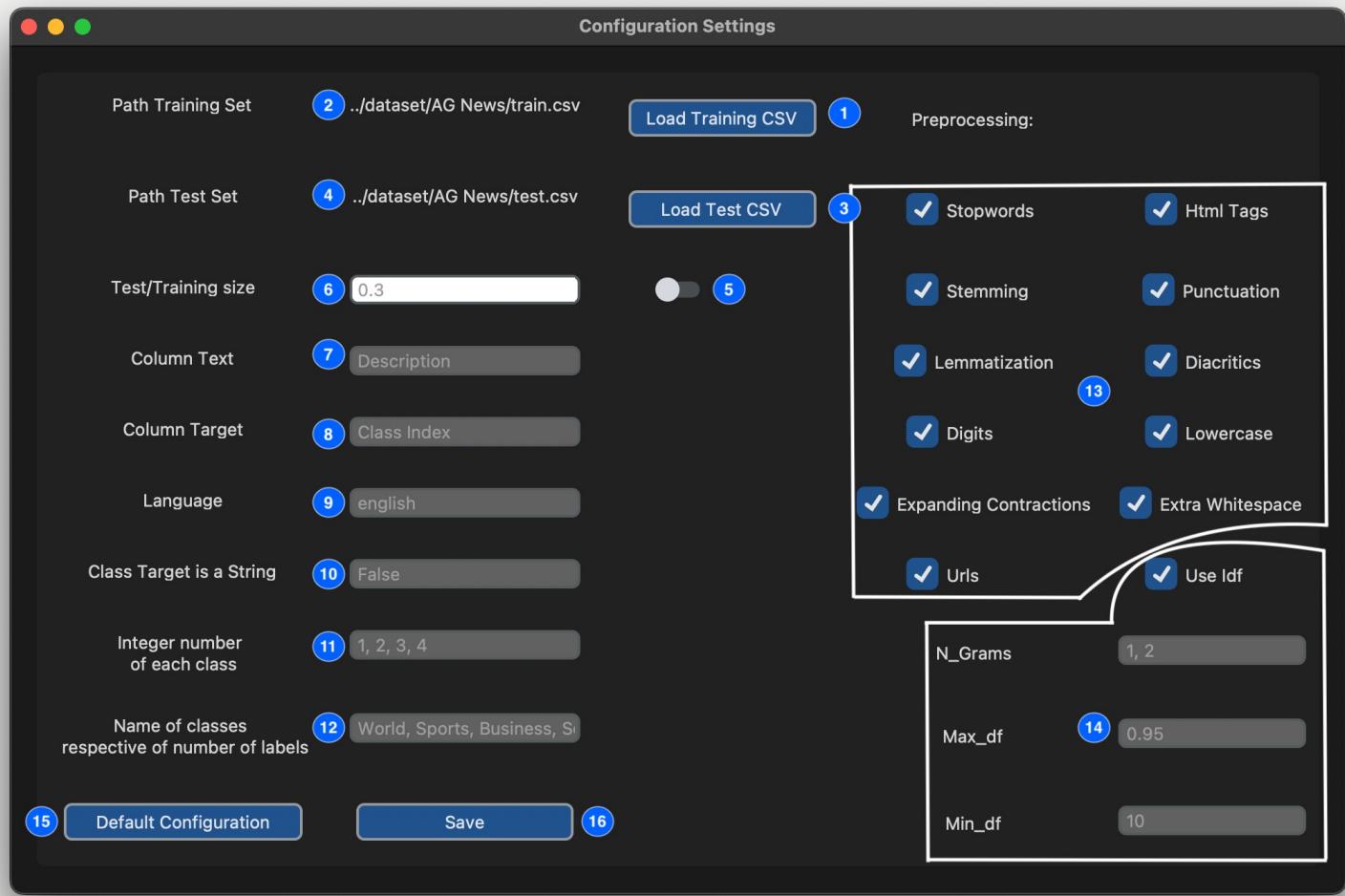
6 Button to change configuration settings

7 Button to see available statistics

8 Button to test models on a different dataset

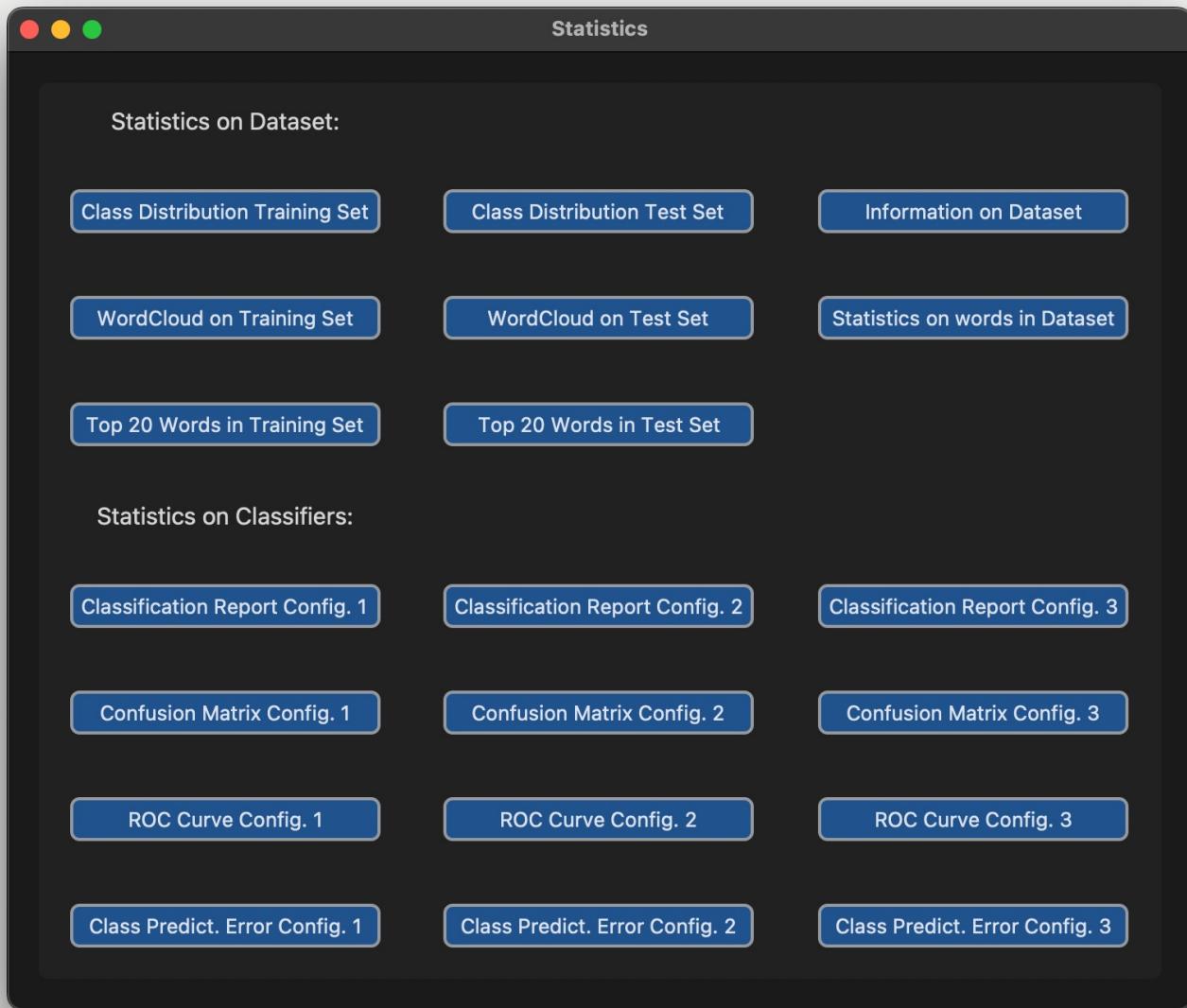
9 Option menu to change appearance theme

CONFIGURATION WINDOW



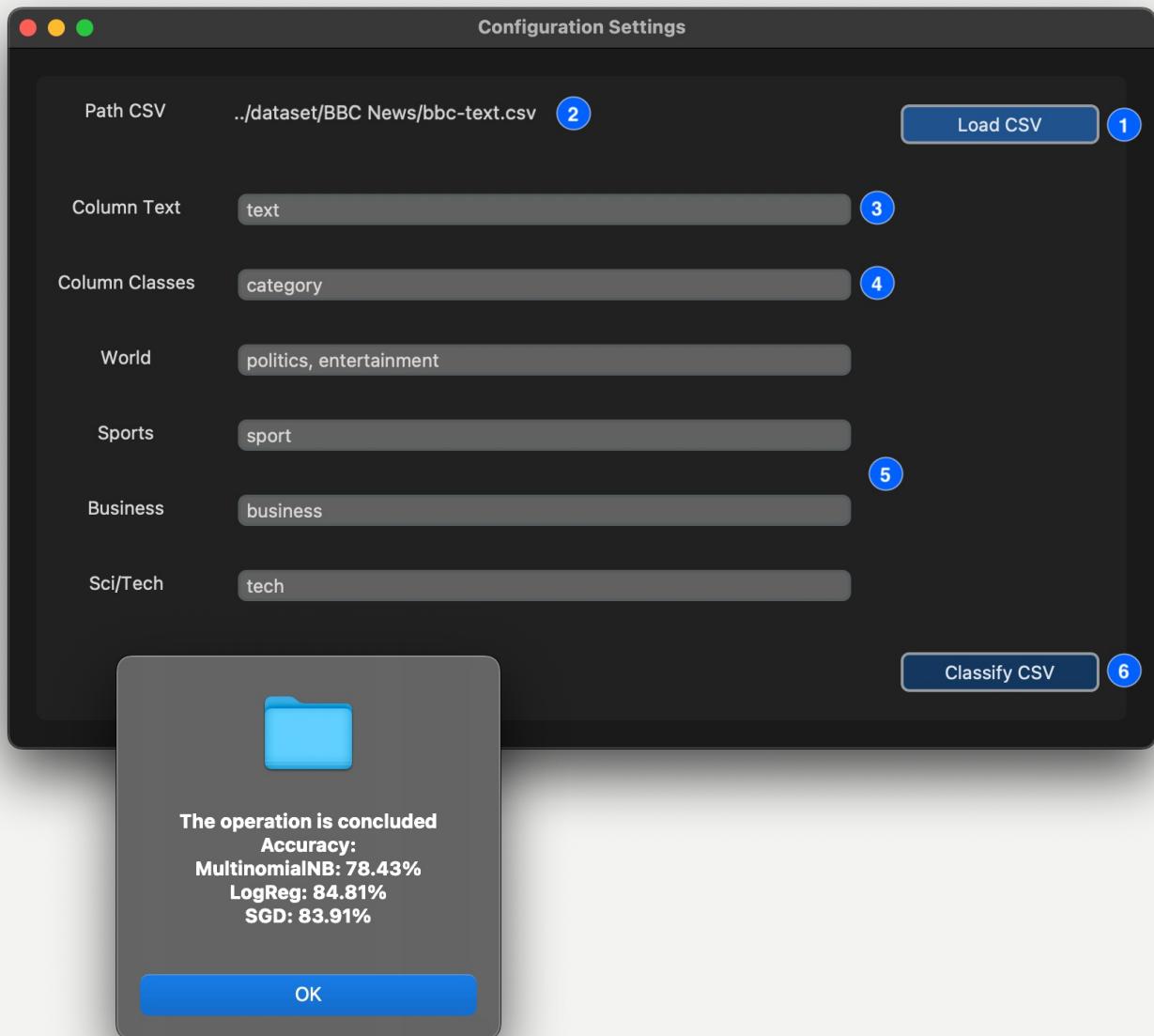
- 1 Button to load training CSV
- 2 Button to load test CSV
- 3 Text displays training CSV path
- 4 Text displays test CSV path
- 5 Switch for one file dataset
- 6 Float test/training size
- 7 CSV column name contains text to classify
- 8 CSV column name contains classes
- 9 Text language
- 10 Boolean
- 11 List of number of classes
- 12 List of classes name
- 13 Preprocess steps
- 14 TdidfVectorizer parameters
- 15 Button to reset configuration
- 16 Button to save configuration

STATISTICS WINDOW



Each button is associate to the corrispective statistic

TEST MODELS WINDOW



- 1 Button to load CSV file
- 2 Text displays path loaded file
- 3 Text column text to classify
- 4 CSV column name contains text to classify
- 5 CSV column name contains classes
- 6 Button to classify CSV file

New csv file cointains three more columns for each classifier:

- 1 Class predicted by specific classifier
- 2 Probability of the class predicted by specific classifier
- 3 0 if class predicted is the same of the right class, 1 if it's not the same