Model Training

Once we have our feature vectors built, we'll try several machine learning classification models in order to find which one performs best on our data. We will try with the following models:

- Baseline Classifier
- Random Forest
- Support Vector Machine
- K Nearest Neighbors
- Multinomial Naïve Bayes
- Multinomial Logistic Regression
- Gradient Boosting

The methodology used to train each model is as follows:

1. First of all, we'll decide which hyperparameters we want to tune.
2. Secondly, we'll define the metric we'll get when measuring the performance of a model. In this case, we'll use the **accuracy**.
3. We'll perform a Randomized Search Cross Validation process in order to find the hyperparameter region in which we get higher values of accuracy.
4. Once we find that region, we'll use a Grid Search Cross Validation process to exhaustively find the best combination of hyperparameters.
5. Once we obtain the best combination of hyperparameters, we'll obtain the accuracy on the training data and the test data, the classification report and the confusion matrix.
6. Finally, we'll calculate the accuracy of a model with default hyperparameters, to see if we have achieved better results by hyperparameter tuning.

We need to be aware of the fact that our dataset only contains 5 categories:

- Business
- Politics
- Sports
- Tech
- Entertainment

So, when we get news articles that don't belong to any of that categories (for example, weather or terrorism news articles), we will surely get a wrong prediction. For this reason we will take into account the conditional probability of belonging to every class and set a lower threshold (i.e. if the 5 conditional probabilities are lower than 65% then the prediction will be 'other'). This probability vector can be obtained in a simple way in some models, but not in other ones. For this reason we will take this into consideration when choosing the model to use.