**Introduction**

The objective of this project is to use a machine learning algorithm to classify text for the purpose of news article recommendation. In this project-specific scenario, there are ten users, each of whom have a unique preferred news topic that they are interested in. The topics are as follows:

* Arts culture entertainment
* Biographies personalities people
* Defence
* Domestic markets
* Forex markets
* Health
* Money markets
* Science and technology
* Share listing
* Sports

These users would like to be recommended up to ten news articles from their preferred topic. It should be noted that some topics may have less than ten articles available, so the algorithm should adjust accordingly.

The objective will be achieved by training an appropriate model before using it to produce the recommendations. Training will be done using a dataset of 9500 news articles. Each article has a corresponding class label belonging to one of the ten topics listed above or labelled as ‘irrelevant’. ‘Irrelevant’ articles do not belong to any of the ten topics and hence are of no interest to the users. Recommendations will be from an unseen dataset of another 500 news articles. The trained model will classify these articles and provide recommendations based on this.

As text pre-processing can be very complex, the articles in the datasets have already been pre-processed. They contain all article words separated by a comma. It is possible that some words occur more than once in the same article, and each distinct word is each treated as a feature.

Hence, we require a machine learning algorithm that is capable of classifying articles based on key words. Classification errors should be reduced as much as possible in order to correctly recommend articles. Thus, the relatively small training set poses an issue as it affects the accuracy of the model produced. Additionally, ‘irrelevant’ articles consist of 49% of the 9500 articles, effectively halving the number of articles related to the topics of interest. This makes it even harder to properly train the model as it is heavily reliant on non-‘irrelevant’ articles.

Due to these limitation, the two models we have chosen to test are regularised linear models with stochastic gradient descent (SGD) learning and Naïve Bayes. SGD has had success with large-scale and sparse machine learning problems. It can scale easily to more than 105 features. Naive Bayes’s main advantages is that good results can be achieved when not much data is available and computational resources are scarce.

**Exploratory Data Analysis**

The class distribution in the training dataset are as follows:

* Arts culture entertainment – 1.2%
* Biographies personalities people – 1.8%
* Defence – 2.7%
* Domestic markets - 1.4%
* Forex markets – 8.9%
* Health – 1.9%
* Money markets – 17.6%
* Science and technology – 0.7%
* Share listing – 2.3%
* Sports – 11.6%

The remaining articles are classes as ‘irrelevant’. We are told that the test set follows the same distribution. Hence, these statistics can be compared with the class distribution obtained by our model on the test set.

**Conclusion**

For this project, we aimed to find an effective machine learning algorithm to classify text for the purpose of recommending up to ten articles across ten different topics.

Summary of method, result, discussion to be done later.