20 Newsgroups - Text Classification Project

Name	Juan Carlos Gutiérrez
Email	juanca.gutierrez@outlook.com
Country	Spain
Specialization	NLP
Repository link	https://github.com/jaycee-ds/20_Newsgroups
Submission date	December 19th 2021

Problem description

We have been given a dataset containing about 20,000 articles that are categorized under one of the following 20 classes:

- alt.atheism,
- comp.graphics,
- comp.os.ms-windows.misc,
- comp.sys.ibm.pc.hardware,
- comp.sys.mac.hardware,
- comp.windows.x,
- misc.forsale,
- rec.autos,
- rec.motorcycles,
- rec.sport.baseball,
- rec.sport.hockey,
- sci.crypt,
- sci.electronics,
- sci.med,
- sci.space,
- soc.religion.christian,
- talk.politics.guns,
- talk.politics.mideast,
- talk.politics.misc,
- talk.religion.misc

This time, we have a text classification problem that must be solved. Given a random article, what if I want to categorize it automatically? A Machine Learning model has been built, so that it helps us to perform this task.

Approach

1. Fetch the dataset.
2. Explore the data.
3. Apply different algorithms and measure their performance.
4. Choose the best model after performing hyperparameter optimization.

The code has been written on the ***20 Newsgroups - Text Classification.ipynb*** file.