20 Newsgroups

Text Classification Machine Learning Problem

Juan Carlos Gutiérrez

NLP specialization

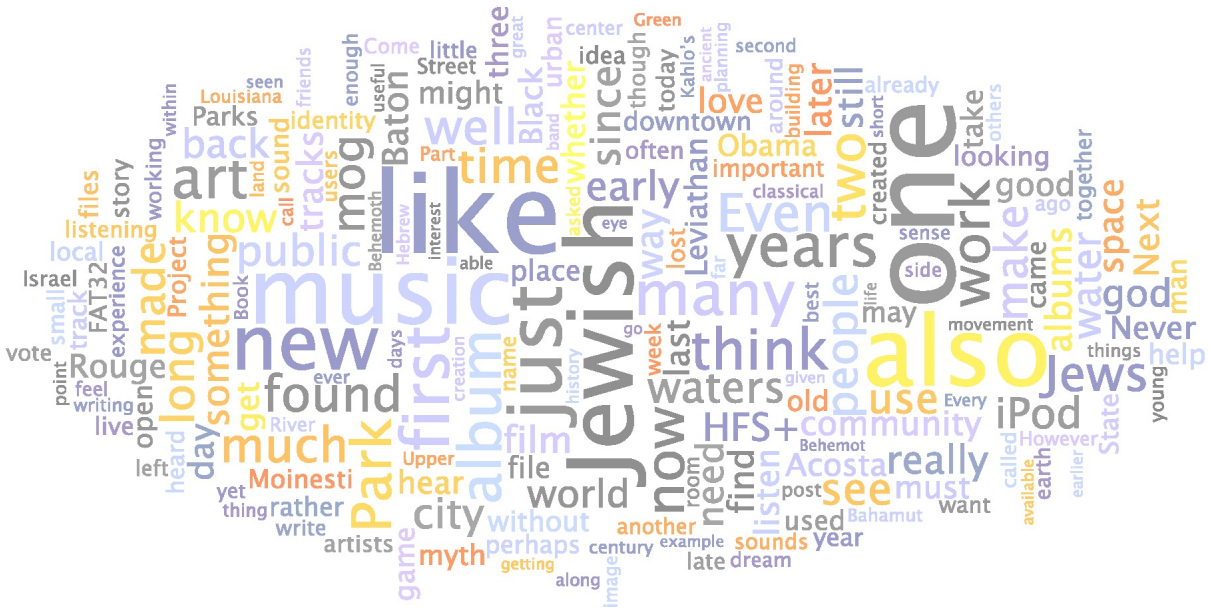
December 19th 2021

Agenda

- Problem Description
- Approach
- Overview of Data
- Models
- Final insights

Problem Description

A Text Classification problem like this is a kind of practice that aims at categorizing a specific text automatically, making use of a Machine Learning model which performs this task for us.



Data Glacier
Your Deep Learning Partner

Approach

Basically, I followed four steps in order to get this task completed:

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

Overview of Data

The dataset contains about 20,000 articles categorized under one of the following buckets:

- ✓ 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

Models

I have applied three different algorithms on the data and measured their performance. Let's summarized this by the accuracy:

- Naïve Bayes -> 77 %
- Support Vector Machines -> 82 %
- XGBoost Classifier -> 75 %

Models

Nevertheless, after performing hyperparameter tuning, I decided to rely on the **Naïve Bayes** model setting parameter alpha to 0.01.

It got 90% accuracy on average after cross-validation over three folds.

Final Insights

Given the results, I recommend to apply a **Naïve Bayes (alpha=0.01)** on the dataset.

Improvements can be made upon deeper hyperparameter optimization, although good results have already been obtained.

An app can also be deployed on the web, so that it can be accessible easily.

Thank You