Classification of political articles.

by Y. Kostrov

Contents

Overview

In this project,
I create a model that will classify political articles collected by me from different websites.

Business Problem

- This project is centered around understanding the political view of the short article that we can find on the internet.
- It is, often, very hard to decide which political inclination the article has.
- The reader can take the text and use the model that comes out of this project to confirm the understanding of the political notion that the user has after reading with the prediction of our model.

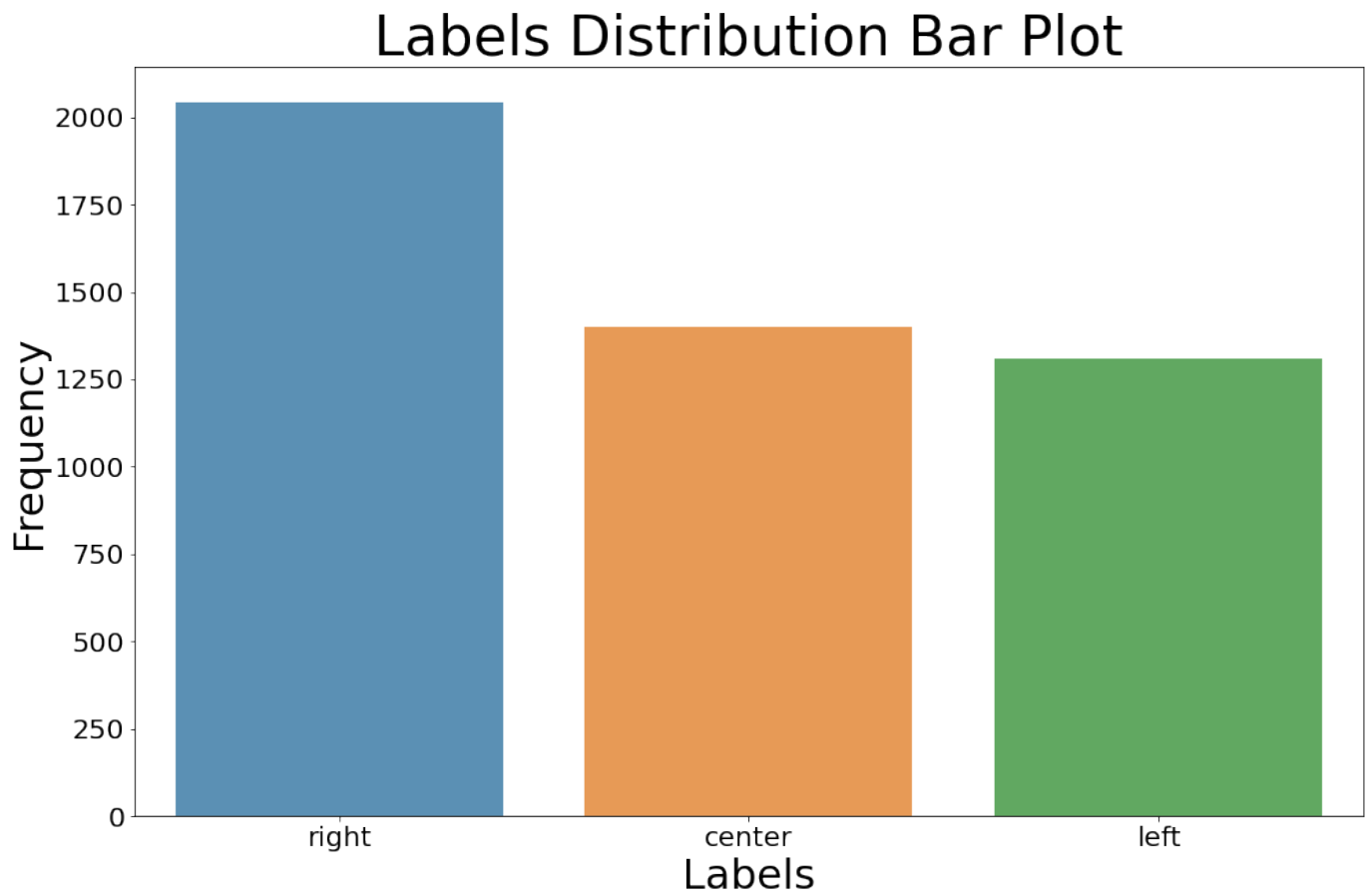
Data Used in the Project

- The data for this project was collected by me from the following websites:
  - www.freebeacon.com,
  - www.americanthinker.com,
  - www.huffpost.com,
  - www.slate.com,
  - news.gallup.com, and
  - www.cbsnews.com.

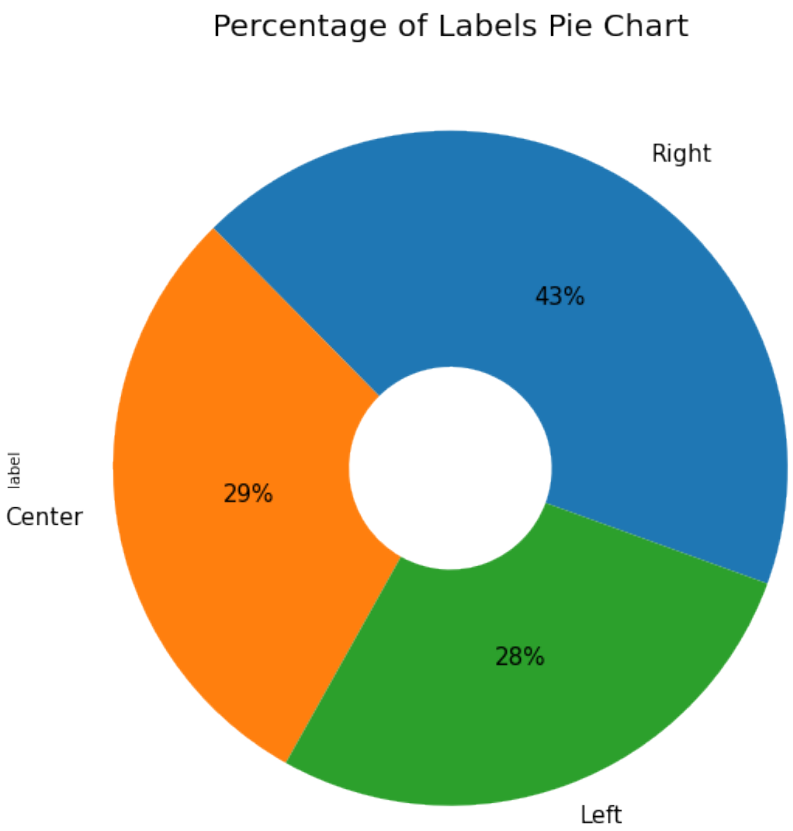
- I put the articles into three category based on the political view of the hosting website.
- The categories are
  - "left",
  - "center"
  - "right".

  .
- The data file has only two columns:
  - "article"
  - "label"

Distribution of Labels by Class



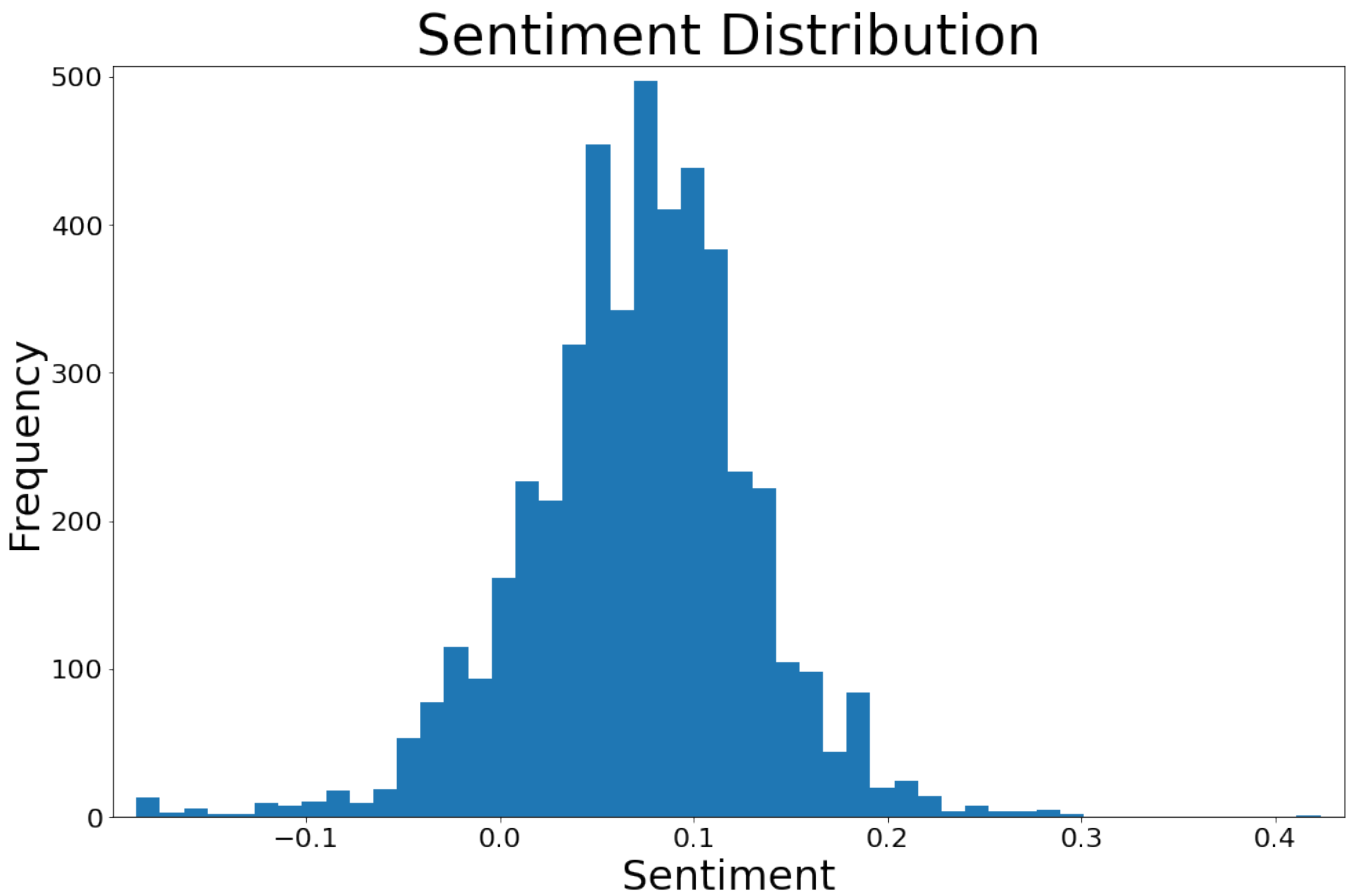
Pie Chart of Labels by Class



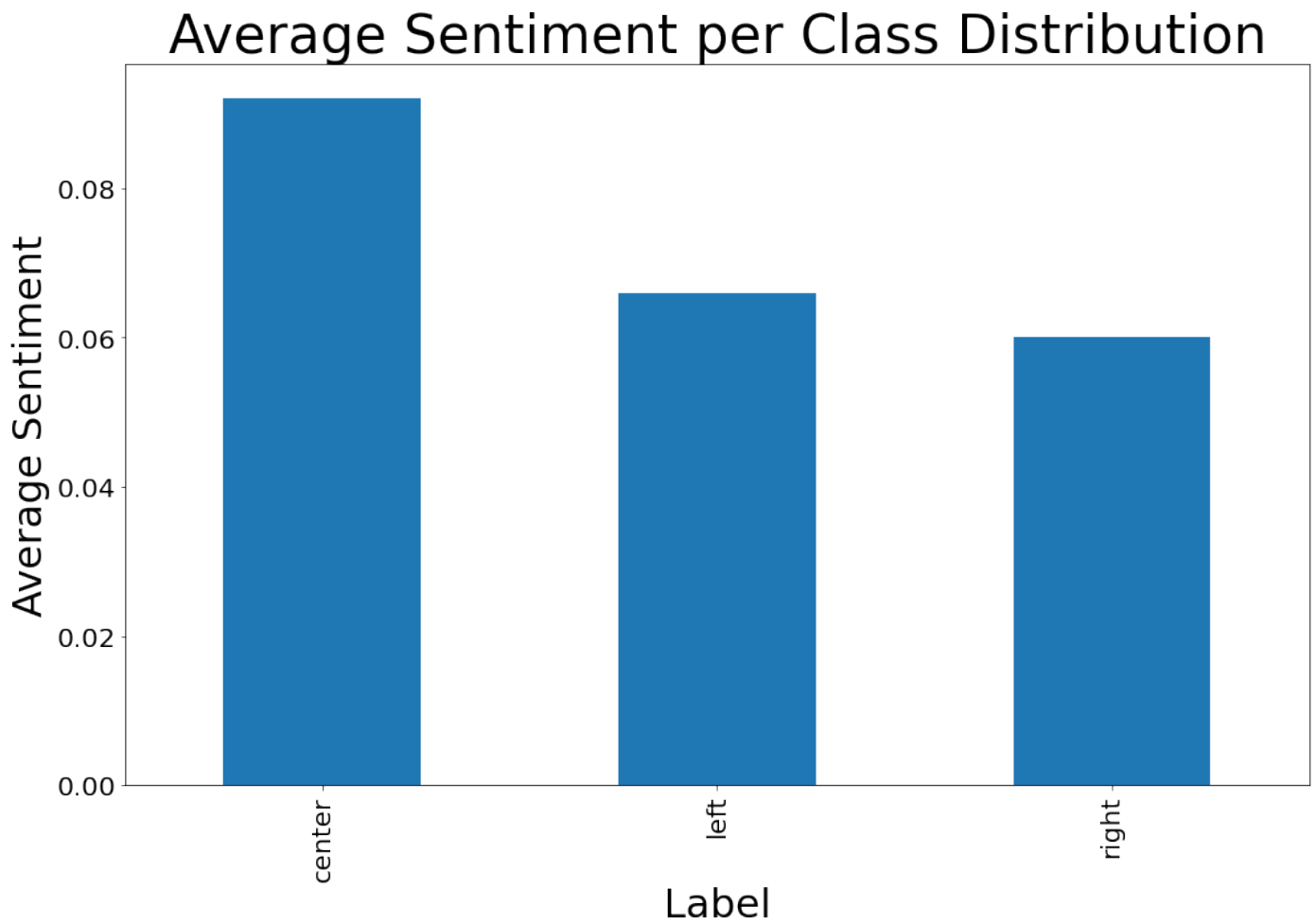
Sentiment analysis

Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. (Wikipedia)

Overall Distribution of Sentiments



Distribution of Average Sentiments by Class



Modeling: Creating Models

- I have used four models: Naive Bayes, Support Vector Machines, Random Forest, and Convolutional Neural Network with Long Short Term Memory Layers with different few different ways to modify text into numerical data suited for machine learning.
- I used accuracy as a primary metric for assessment of models.

Explanation of Accuracy

- There is another metric we will use is called "accuracy".
- Accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{All Predictions}}$$

Explanation of Accuracy

- Accuracy is the number of correctly predicted data points out of all the data points.

How Well Models Performed

- The best performance was achieved with the Support Vector Machines classifier at 98.7% of accuracy.

Business Suggestion

Based on my analysis,

- I suggest to use Support Vector Machines model for the prediction of the political inclination of the article.
- The model can help internet users who are not sure about the flavor of the political news to verify what kind of an article they are reading.

THE END
THANK YOU!