Article Classification with Centroid Classification

Group 1

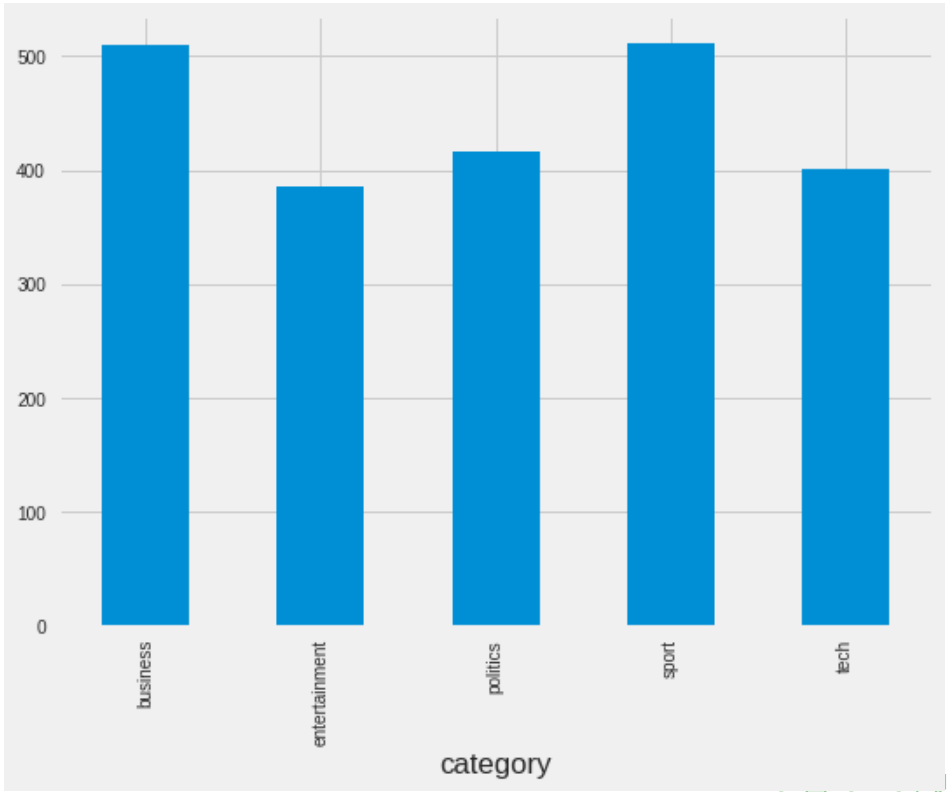
June 19, 2024

Project Objectives

- The goal of this project is classify document per category on given dataset.
- Document Embedding Model: Tf-Idf Vectorizer
- Classification Model: Centroid Classification

Dataset and Pre-liminary Analysis

- BBC ML Dataset
- Consist of 2225 articles, with categories: 'business', 'entertainment', 'politics', 'sport', 'tech'
- Dataset was changed from Wikipedia to BBC dataset due to complication in model training and dataset processing.



Data Preprocessing

- Category-Token Chi2 test

Category	1st	2nd	3rd
Business	oil	bank	growth
Entertainment	award	singer	awards
Politics	tory	blair	party
Sport	win	injury	coach
Tech	digital	computer	software

Tf-Idf Embedding

- We use tf-idf scoring to encode the documents.
- The token value is calculated as follows:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

$$idf(t) = \log\left(\frac{N}{n_t}\right) \tag{2}$$

$$tfidf(t,d) = tf(t,d) \times idf(t) \tag{3}$$

Tf-Idf Embedding

- Prior to embedding, unique words(unigrams and bigrams) are extracted from the dataset to be used as tokens. Each token is associated with token index which will be used as vector index.
- For each document, the token value(vector index) in the vector is given by tf-idf scoring of document.

Document Vectorization and Clustering

- With the documents embedded, we now calculate the centroid and variance of each categories.
- Centroid is the main metric for the cluster, and variance is calculated to use gaussian distribution for probability distribution.
- No dimension reduction is applied to minimize information loss.
- We evaluate the model on normalized dataset and non-normalized dataset. For normalized dataset, we use distance matric $\frac{1}{\cos(\theta)}$, where $\cos(\theta)$ is the cosine similarity, and for non-normalized dataset, we use euclidean distance.

Centroid Classification

- Centroid classification is a simple classification method that uses the centroid and distance metric to classify the document.
- Our project uses method: Nearest Centroid, Gaussian Probability Distribution, and Logistic function.
- The category prediction is generated by calculating highest score of probability distribution, or lowest distance from the centroid.

Nearest Centroid

- The simplest centroid classification method, where the document is classified based on the nearest centroid.
- The distance metric is calculated using the distance metric, and the document is classified based on the nearest centroid.
- The centroid is calculated as the mean of the document vectors in the cluster.

Gaussian Probability Distribution

- For each unique category, we construct a probability distribution using the centroid of the cluster and distance metric.
- Assuming central-limit theorem, we use a normal distribution to generate the probability distribution.

$$p(d) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d)^2}{2\sigma^2}}$$

Where d is the distance from the centroid, and σ is the variance of the cluster.

Logistic Function

- Use Logistic Function to smooth the distance metric.

$$p(d) = \frac{1}{1 + e^{-\alpha d}}$$

Where d is the distance from the centroid. We use α as a hyperparameter to tune the distribution, and it's set to 0.5.

Multi-Category Classification

- For multi-category classification, we use the probability distribution to predict the category of the document.
- We can tune the probability threshhold for accuracy using the cost function:

$$J(C_d, C'_d) = \frac{\lambda_1 |C'_d \setminus C_d| + \lambda_2 |C_d \setminus C'_d|}{|C_d \cup C'_d|}$$

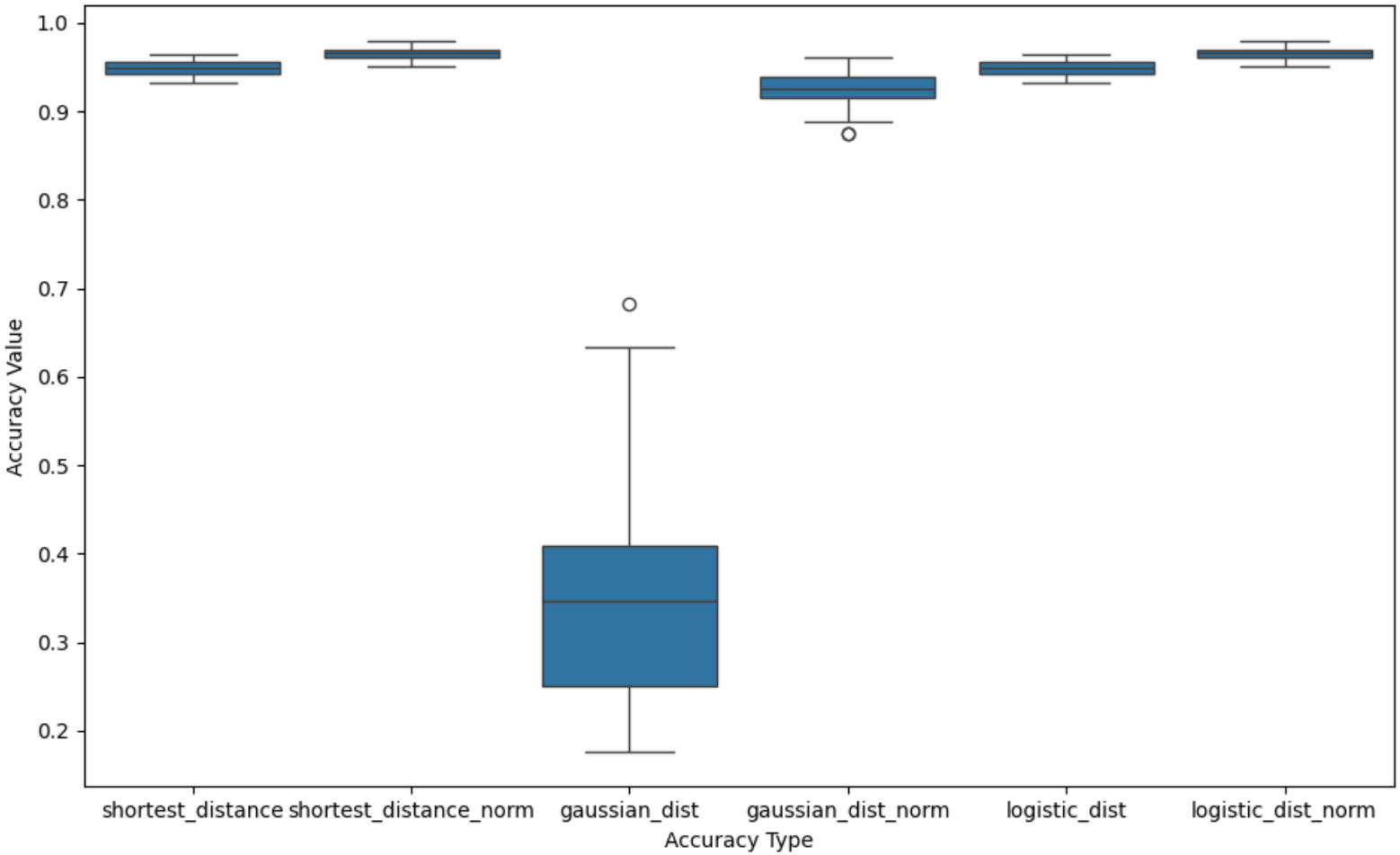
  where C_d is document categories set, and C'_d is predicted categories set.
- Starting from threshold 0, we increase the threshold by a learning rate value until the cost function is minimized.
- We generate a candidate category for both pmf value using trained threshold, and choose candidate of which have same candidate for Gaussian and Logistic.

Evaluation

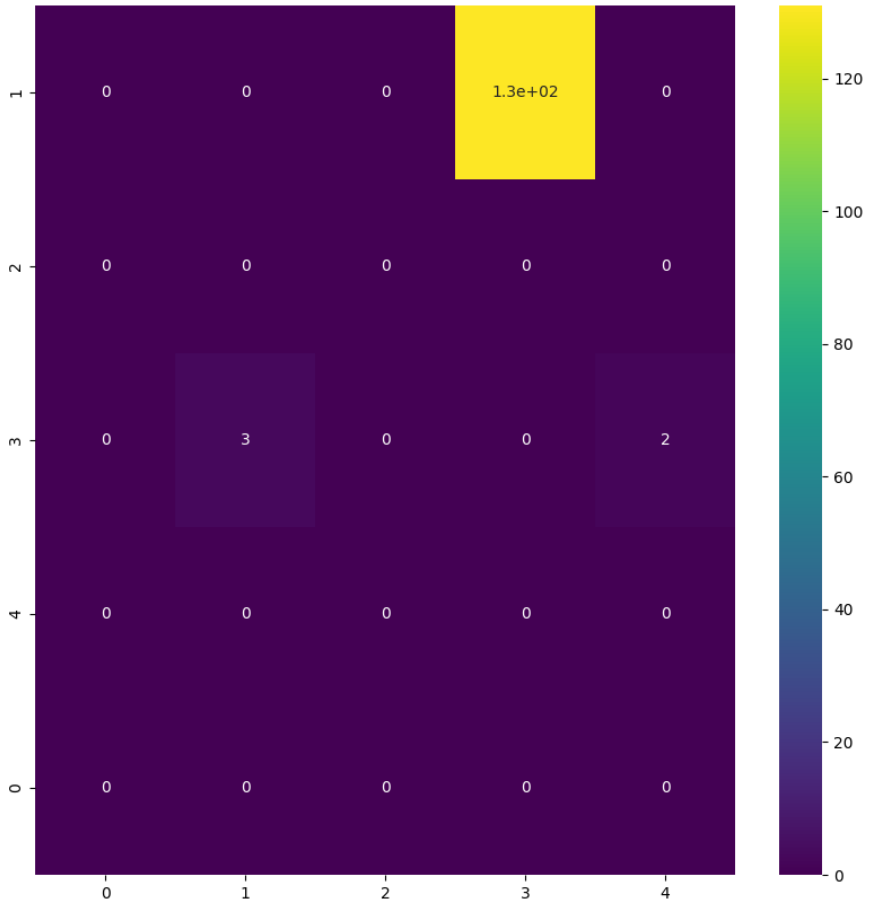
- Model Accuracy

Category	Accuracy
Closest Centroid	0.95
Closest Centroid Normalized	0.97
Gaussian	0.34
Gaussian Normalized	0.93
Logistic	0.95
Logistic Normalized	0.97

Evaluation


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Accuracy Value
shortest_distance
shortest_distance_norm
gaussian_dist
gaussian_dist_norm
logistic_dist
logistic_dist_norm
Accuracy Type

Multiple Category Classification



Multiple Category Classification

- Sports article were most associated with entertainment, while entertainment article was candidate for only 3 sports. The 3 documents commonly include description of winning an award or similar context
- 1 tech documents were incorrectly labled as sports, which included description of video game metal slug.

Entertainment with Candidate Sports

Ray's success on DVD outstripped its $74m (£40m) US box office total, earning more than $40m (£22m) on the first day of the DVD's release alone. Ray has been nominated in six Oscar categories including best film and best actor for Jamie Foxx. The film recounts the life of blues singer Ray Charles, ...

Miscategorized Tech Article

Like some drill sergeant from the past, Metal Slug 3 is a wake-up call to today's gamers molly-coddled with slick visuals and fancy trimmings. With its hand-animated sprites and 2D side-scrolling, this was even considered retro when released in arcades four years ago. But a more frantic shooter you will not find at the end of your joypad this year. And yes, that includes Halo 2. Simply choose your grunt and wade through five 2D side-scrolling levels of the most hectic video game blasting you will ever encounter. It is also the toughest game you are likely to play, as hordes of enemies and few lives pile the pressure on...