**Media Coverage Tracker** (News Classification)

Scrape daily news articles about coal, energy, or CCL, and classify them as “Positive PR”, “Negative PR”, or “Neutral”. Why useful: PR can track their brand’s media coverage daily. Beginner-friendly tools: Python + Scikit-learn (Naive Bayes classifier).

Steps:

1. Created a dummy dataset
2. Vectorize text
3. Detect Patterns using Naïve Bayes Classifier
4. Train using Training data
5. Test on new articles
6. Tried enhancing using SVM.

Why Naïve Bayes?

Naive Bayes is one of the most popular algorithms for **text classification** (like spam detection, sentiment analysis, or PR tone classification) because:

* Text data naturally fits its assumptions.
* It performs surprisingly well even on small datasets.
* It’s very fast to train and predict.

Naive Bayes uses **probabilities** based on how often words appear in each category.  
For example:

* If words like *“award,” “sustainable,” “growth”* appear often in *Positive PR* articles,  
  → the model learns that these words are strong indicators of *Positive PR*.
* If words like *“pollution,” “criticism,” “lawsuit”* appear more in *Negative PR* ones,  
  → it learns those are negative indicators.

When a **new article** appears, it checks which label’s word pattern it matches most closely.

Why do we vectorize text?

**Naive Bayes (and all ML models) work with numbers — not words**

Machine learning algorithms, including Naive Bayes, **can’t understand raw text** like this:

"Coal company invests in renewable energy projects"

They only work with **numerical features** (matrices of numbers).  
So we must **convert text → numbers** while keeping the meaning or patterns of words.

That’s exactly what **vectorization** does.

How to vectorize:

We turn each article (text) into a **vector** — a list of numbers that represents the text.

Example using **CountVectorizer**:

For eg; Articles:

1. Coal company invests in renewable energy projects

2. Coal mining causes environmental damage

After vectorization:

| Word | coal | company | invests | renewable | energy | projects | mining | causes | damage | environmental |  
|----------------------|------|----------|----------|------------|---------|-----------|---------|---------|----------------|  
| Article 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |  
| Article 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

Each row = article  
Each column = word count

The **Multinomial Naive Bayes** model uses these counts to calculate probabilities:

* How often does each word appear in *Positive PR*, *Negative PR*, or *Neutral* articles?
* When a new article arrives, it checks which class’s word patterns match best.

So if the word “award” appears often in *Positive PR*, the model increases the probability that a new article containing “award” is *Positive PR*.

* **CountVectorizer** → counts how often each word appears.
* **TF-IDF Vectorizer** → gives higher importance to unique words and downweights common ones like “the” or “is”.

In this project, we used **CountVectorizer** (simpler and good for small datasets).

Classification Report:

precision recall f1-score support

Negative PR 0.00 0.00 0.00 2.0

Positive PR 0.00 0.00 0.00 0.0

accuracy 0.00 2.0

macro avg 0.00 0.00 0.00 2.0

weighted avg 0.00 0.00 0.00 2.0

Confusion Matrix:

[[0 2]

[0 0]]

After testing on new articles:

Article: "CCL invests in new wind energy project" => Predicted Label: Positive PR

Article: "Coal plant faces lawsuit for pollution" => Predicted Label: Negative PR

Article: "Energy market remains stable this quarter" => Predicted Label: Positive PR

**SVM Results**: After TF-IDF Vectorization (gives better separation between common and rare words)

Classification Report:

precision recall f1-score support

Negative PR 0.00 0.00 0.00 2

Positive PR 0.33 1.00 0.50 1

accuracy 0.33 3

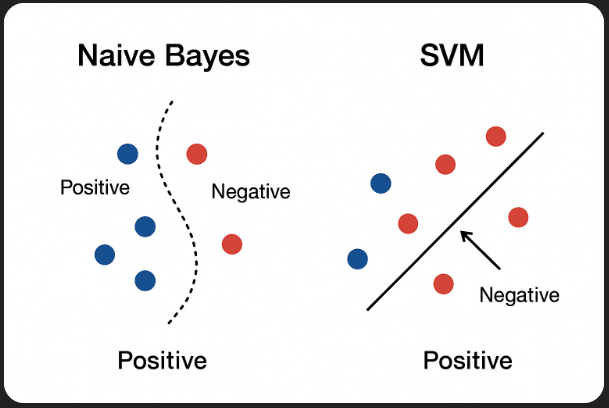
macro avg 0.17 0.50 0.25 3

weighted avg 0.11 0.33 0.17 3

Confusion Matrix:

[[0 2]

[0 1]]



After testing on new articles:

Article: "CCL invests in new wind energy project" => Predicted Label: Positive PR

Article: "Coal plant faces lawsuit for pollution" => Predicted Label: Negative PR

Article: "Energy market remains stable this quarter" => Predicted Label: Positive PR

Colab Link:

<https://colab.research.google.com/drive/1YzUtUwGZaV5Nm_gMxhzx0SSIPQgHY-64#scrollTo=PuWva5nupisl>