📰 News Article Classification Project Summary

🎯 Objective

To develop a machine learning model capable of classifying news articles into predefined categories (e.g., politics, entertainment, wellness) using text data from headlines and descriptions.

📥 1. Data Collection & Exploration

- **Source:** news.csv file with 50,000 articles.
- **Fields used:** headline, short_description, and category.
- Combined headline and short_description into a new column text.
- Focused on top 7 categories to reduce label noise.

🧹 2. Data Preprocessing

- Lowercased all text.
- Removed URLs, HTML tags, punctuation, and stopwords using nltk.
- Applied word tokenization and cleaned whitespaces.
- Created a clean_text column with preprocessed content.

📊 3. Exploratory Data Analysis (EDA)

- Plotted the distribution of articles across categories.
- Extracted top keywords per category to identify themes using word frequencies.

🔍 4. Feature Engineering

- **Bag of Words (BoW):** Used CountVectorizer with 5,000 most frequent words.
- **TF-IDF:** Used TfidfVectorizer (up to bigrams, 8,000 features).
- **Word2Vec:** Created 100-dimensional vectors using gensim's Word2Vec.

🧪 5. Model Development

Used TF-IDF features (X) for model input and encoded categories for output (y).

🧠 Models Applied:

Model	Accuracy	Notes
Logistic Regression	**83.6%**	Best performer; robust with tuning (C, solver)
Support Vector Machine (Linear SVC)	83.7%	Comparable to LR with strong generalization.
Naive Bayes	82.4%	Fast, simple, effective with TF-IDF.

🤝 6. Ensemble Learning

- Used **Voting Classifier** with LR, SVM, and NB.
- **Cross-Validation Accuracy:** 83.8%, slightly better than any individual model.

📈 7. Evaluation Metrics

- Used:
  - **Accuracy**
  - **Precision, Recall, F1-score**
  - **Confusion Matrix**
- All models showed **balanced performance** across categories.
- Ensemble slightly improved stability and overall accuracy.

🧠 Key Insights

- Preprocessing and TF-IDF played a crucial role in boosting model accuracy.
- Logistic Regression was interpretable and highly effective.
- Combining models in an ensemble proved marginally more accurate and robust.

📽️ Presentation

A video walkthrough was also created to explain the workflow:

https://drive.google.com/file/d/10GjSjZYA67LIcRuqYD9slTI1K-Z9Rc2l/view?usp=sharing