**News Article Topic Classifier**

A Final Project

Presented to the

Faculty College of Computer Studies

Laguna State Polytechnic University-San Pablo City Campus

San Pablo City, Laguna

\_\_\_\_\_\_\_\_\_\_

In Partial Fulfillment

of the Requirements for the Degree of

Bachelor of Science in Computer Science

Specialized in Intelligent System

\_\_\_\_\_\_\_\_\_\_

By:

**PRINCE MAYNARD M. MASILAC**

**ANDREI B. PIRA**

**GEORGE JOSEPH M. SIA**

May 2025

# **Abstract**

This project explores the application of Natural Language Processing (NLP) for classifying news articles into predefined categories. Using the AG News Classification Dataset, we developed a text classification model that leverages traditional NLP techniques and a Logistic Regression algorithm. The pipeline includes text preprocessing, TF-IDF vectorization, and supervised learning to predict news categories such as World, Sports, Business, and Sci/Tech. The model provides real-time predictions based on user input and demonstrates the effectiveness of classical machine learning models in NLP tasks.

# **Introduction and Background**

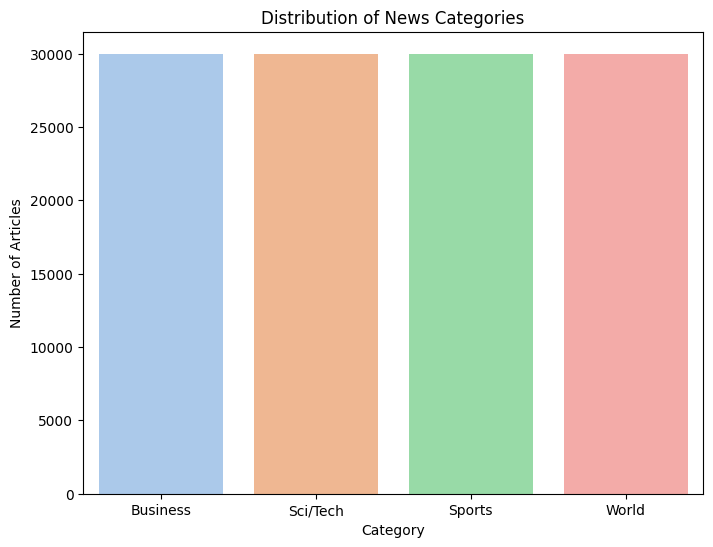
Text classification is a foundational task in Natural Language Processing (NLP), involving the assignment of predefined categories to textual data. It is widely applied in domains such as spam detection, sentiment analysis, and customer feedback analysis, as well as in media analytics, particularly news classification (Kowsari et al., 2019). With the exponential growth of digital news content, there is a growing demand for intelligent systems that can automatically organize and categorize articles to enhance searchability, personalization, and information retrieval.

In this project, we tackle the task of classifying news headlines into four high-level categories: World, Sports, Business, and Science/Technology. While recent advancements in NLP often rely on complex deep learning architectures like BERT or GPT (Devlin et al., 2019), our project emphasizes the value of interpretable and computationally efficient traditional methods. Specifically, we utilize a Logistic Regression classifier within a classical NLP pipeline consisting of text normalization, tokenization, stopword removal, and TF-IDF vectorization.

Logistic Regression has long served as a strong baseline for text classification, especially when coupled with well-crafted textual features (Rennie et al., 2003). This project not only demonstrates the real-world utility of traditional NLP techniques but also shows how thoughtful preprocessing can yield strong results even without deep neural networks.

# **Dataset Description and Preprocessing**

The dataset used in this project is the AG News Classification Dataset from Kaggle, a well-known benchmark for topic classification tasks in NLP. It consists of 120,000 training samples and 7,600 testing samples, each comprising a short news title and a corresponding category label. The labels are distributed evenly across four categories: World, Sports, Business, and Sci/Tech. The balanced nature of the dataset in **Figure 1** makes it ideal for supervised learning algorithms, ensuring that the model is not biased toward any single category.



**Figure 1: Class Distribution**

Before feeding the text data into the machine learning model, several preprocessing steps were applied to clean and standardize the input. First, all text was converted to lowercase to eliminate case-based variance. Then, punctuation marks and non-alphanumeric characters were removed using regular expressions. The text was tokenized into individual words, and common stopwords (such as "the", "is", and "and") were filtered out using the NLTK library. These steps help reduce noise and focus on informative words. Finally, the processed text was transformed into numerical feature vectors using Term Frequency-Inverse Document Frequency (TF-IDF), which quantifies the importance of each word in the context of the entire dataset. This final representation enables the Logistic Regression model to effectively learn patterns in word usage that are indicative of different news topics.

# **Final Output and Analysis**

To better visualize the final output, a simple console interface in **Figure 2** allows the user to input a news title. Upon entering a news headline, the model processes the input using the trained model and returns a predicted category such as World, Sports, Business, or Sci/Tech. This functionality demonstrates the model's ability to generalize beyond the training data and classify real-world text inputs accurately.



**Figure 2: Final Output**

To evaluate the model's performance, we utilized standard classification metrics: accuracy, precision, recall, and F1-score. On the AG News test set, the model achieved an accuracy of approximately 90%. The confusion matrix revealed that the classifier was particularly accurate in distinguishing *Sports* and *Business* news, likely due to the presence of domain-specific keywords (“goal,” “team,” “market,” “stocks”, etc...). However, some confusion was observed between *World* and *Sci/Tech* categories, which is expected given the occasional overlap in terminology and subject matter. We also conducted a few qualitative tests as we can see in **Figure 2,3 and 4** by manually entering various sample headlines to assess real-world usability.

  
**Figure 3: Testing**



**Figure 4: Testing**

****

**Figure 5: Testing**

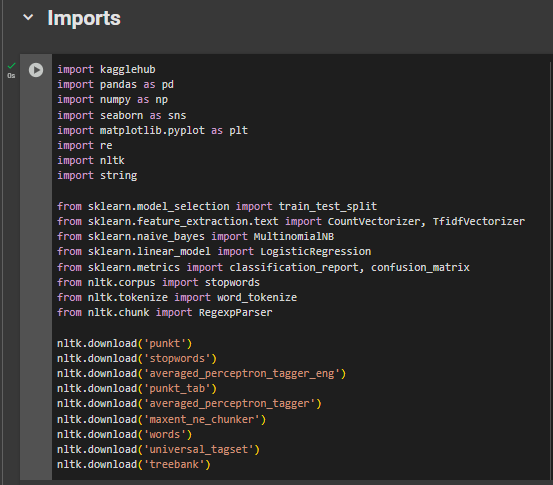
These results illustrate that the model can correctly interpret the semantic cues in short headlines and associate them with appropriate categories. Despite its simplicity, the logistic regression approach proved highly interpretable, efficient, and capable of producing reliable predictions without the need for resource-heavy deep learning architectures.

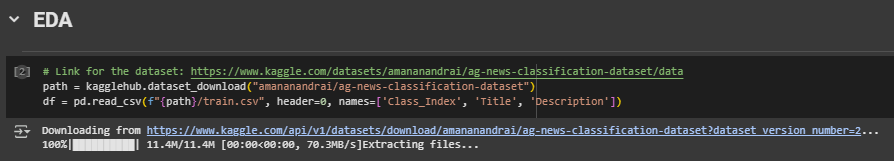
# **Conclusion and Possible Future Work**

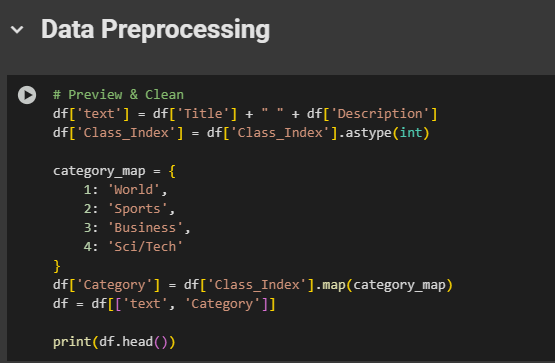
In this project, we successfully developed a news article topic classifier using a logistic regression model trained on the AG News dataset. By applying core NLP techniques such as text normalization, tokenization, stop-word removal, TF-IDF vectorization, and text classification, we demonstrated the practical application of machine learning in categorizing news headlines into distinct categories. The model achieved high accuracy and showed strong performance in real-time predictions, confirming its effectiveness for short-text classification tasks. One of the strengths of this approach is its interpretability and computational efficiency, making it suitable for educational and lightweight deployment scenarios.

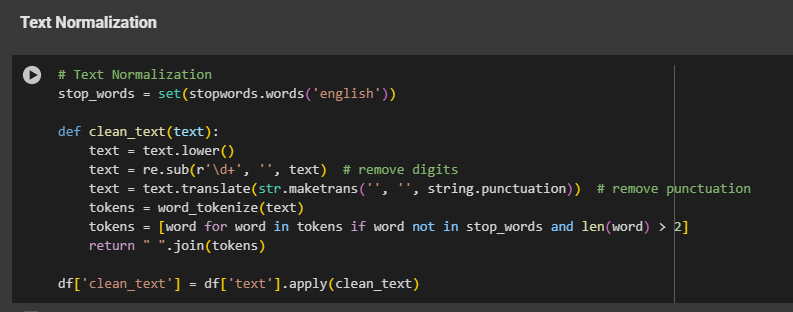
For future work, the model could be enhanced by integrating more advanced NLP techniques such as word embeddings (e.g., Word2Vec, GloVe) or deep learning models like BERT for better contextual understanding. Additionally, extending the dataset to include longer article content or more granular subcategories could improve classification depth. Finally, incorporating a user-friendly interface or a web-based application would increase the system's accessibility and real-world utility, especially for media platforms or news monitoring tools.

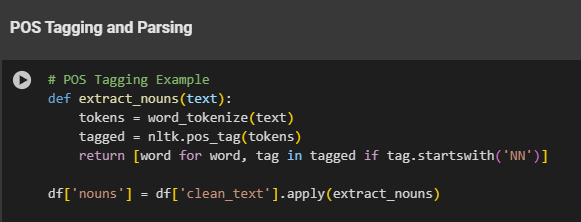
# **Appendix**

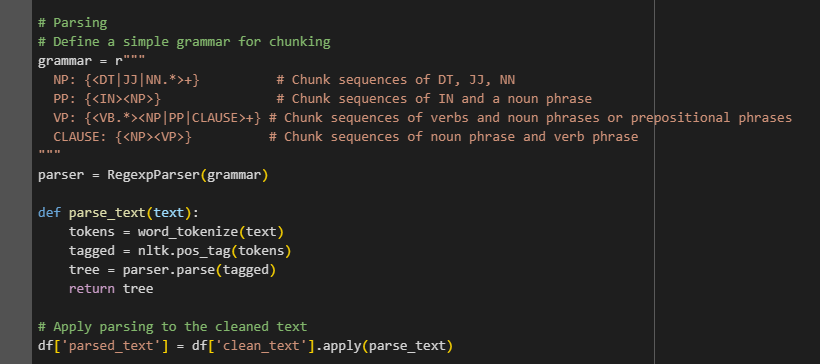


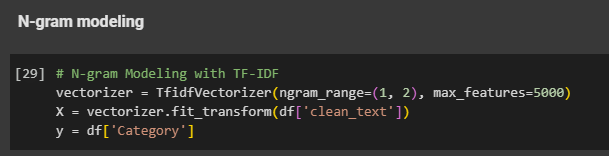


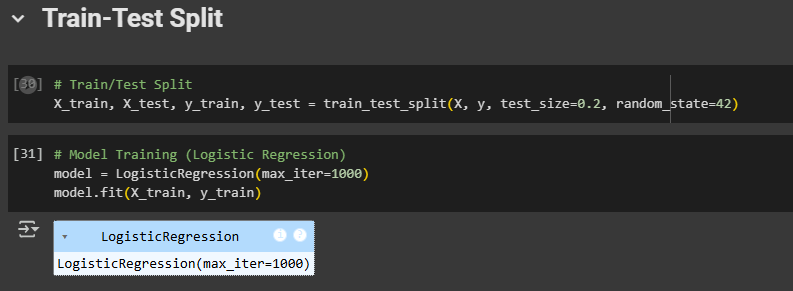


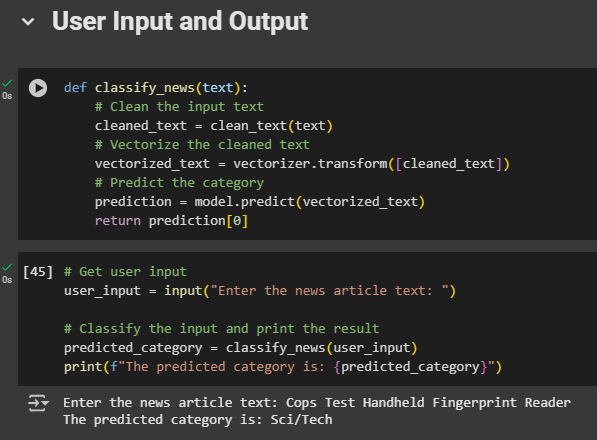












# **References**

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. (2019). *Text Classification Algorithms: A Survey*. Information, 10(4), 150.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. (2003). *Tackling the poor assumptions of naive bayes text classifiers*. In ICML.