Topic Modelling Using AG News Dataset

EAS 503LEC JLI: Prog DB Data Sci

PROFESSOR:
Jianzhen Liu

PREPARED BY :
Sai Swapnesh Pahi
Akanksha Kale

Introduction:

In an era of vast digital content, recommendation systems play a crucial role in helping users navigate vast amounts of content. However, many articles lack explicit genre labels, complicating effective classification and recommendations. This report explores the application of topic modeling techniques to infer hidden genres from unclassified articles, aiming to enhance the accuracy of recommendation systems.

Objectives:

The project aims to develop a machine learning system that automatically classifies news articles into relevant genres such as World, Sports, Business, and Science/Technology, thereby streamlining the content organization. This system will serve as the foundation for recommendation engines, allowing them to suggest articles tailored to user preferences and enhancing the overall user experience on digital platforms. Additionally, the project seeks to create a scalable solution that utilizes text embeddings and efficient classification algorithms, ensuring the ability to handle large datasets and diverse text content effectively.

Dataset Description:

The AG News dataset is a widely used resource for text classification, containing 120,000 training samples and 7,600 test samples of English-language news articles. Each article is categorized into one of four classes: World, Sports, Business, and Science/Technology. Developed by Xiang Zhang, this dataset is commonly employed in machine learning and natural language processing research, particularly for tasks such as news categorization, topic modeling, and the development of recommendation systems.

Dataset relevance to the topic:

The AG News dataset is well-suited for topic modeling in news recommendation systems, categorizing articles into four genres: World, Sports, Business, and Science/Technology. This alignment with typical news genres supports effective training and evaluation of machine learning models, enabling accurate genre predictions even without explicit metadata labels. This diversity ensures robust analysis, contributing to the development of systems that assign genre-based recommendations.

Key Variables and Their Significance:

- Text: The main feature containing the news article content, used to extract and analyze topic-related embeddings.
- Label: The category assigned to each article (World, Sports, Business, Science/Technology). This variable acts as the reference for supervised learning tasks and helps evaluate the accuracy of the topic modeling method, allowing for performance assessment and improvement of the model.

Anticipated Challenges:

- Extracting Text Embeddings: Generating 768-dimensional embeddings from DistilBERT is computationally intensive and may face memory and processing bottlenecks with large datasets.
- Class Imbalance: Any imbalance in the dataset could affect the model's ability to generalize across all genres.
- Handling Text Length Variations: Variations in article length may impact model performance, with shorter texts lacking context and longer ones introducing noise.

Data Processing Workflow:

- Dataset Import and Setup: Loaded the AG News dataset and configured DistilBERT with GPU processing to generate 768-dimensional embeddings.
- Database Design and Structure: Created a database with two tables: raw_data for storing text and labels, and embeddings_data for storing processed embeddings linked to raw_data.
- Concurrent Processing: Processed text data in batches of 512 using a T4 GPU, ensuring efficient embedding extraction.
- Dimensionality Reduction and Visualization: Used T-SNE to reduce the 768-dimensional embeddings to 2D for visualization, selecting 1,000 records for an interactive scatter plot in a Streamlit app to explore data clusters.

Critical Steps in SQL Database Design:

- Data Structure Definition and Table Creation
- Created two tables: raw_data (stores text, labels, and unique IDs) and embeddings_data (stores processed embeddings, with references to raw_data).
- Designed raw_data to store id (Primary Key), text, and label; embeddings_data to store id (Primary Key), raw_data_id (Foreign Key), embedding (BLOB), and label.
- Database Initialization and Data Population: Established an SQLite connection and initialized the cursor. Inserted data into raw_data using SQL INSERT statements, mapping IDs, text, and labels. Populated embeddings_data with extracted text embeddings, linking them to raw_data_id and labels.
- Batch Processing and Efficient Embedding Extraction: Utilized concurrent batch processing with GPU acceleration to efficiently handle embedding extraction.
- Data Retrieval and Integration for Visualization: Executed SQL queries to retrieve data from the tables. Merged 1,000 sampled records from embeddings_data with raw_data for dimensionality reduction and scatter plot visualization.

Analysis Process Overview:

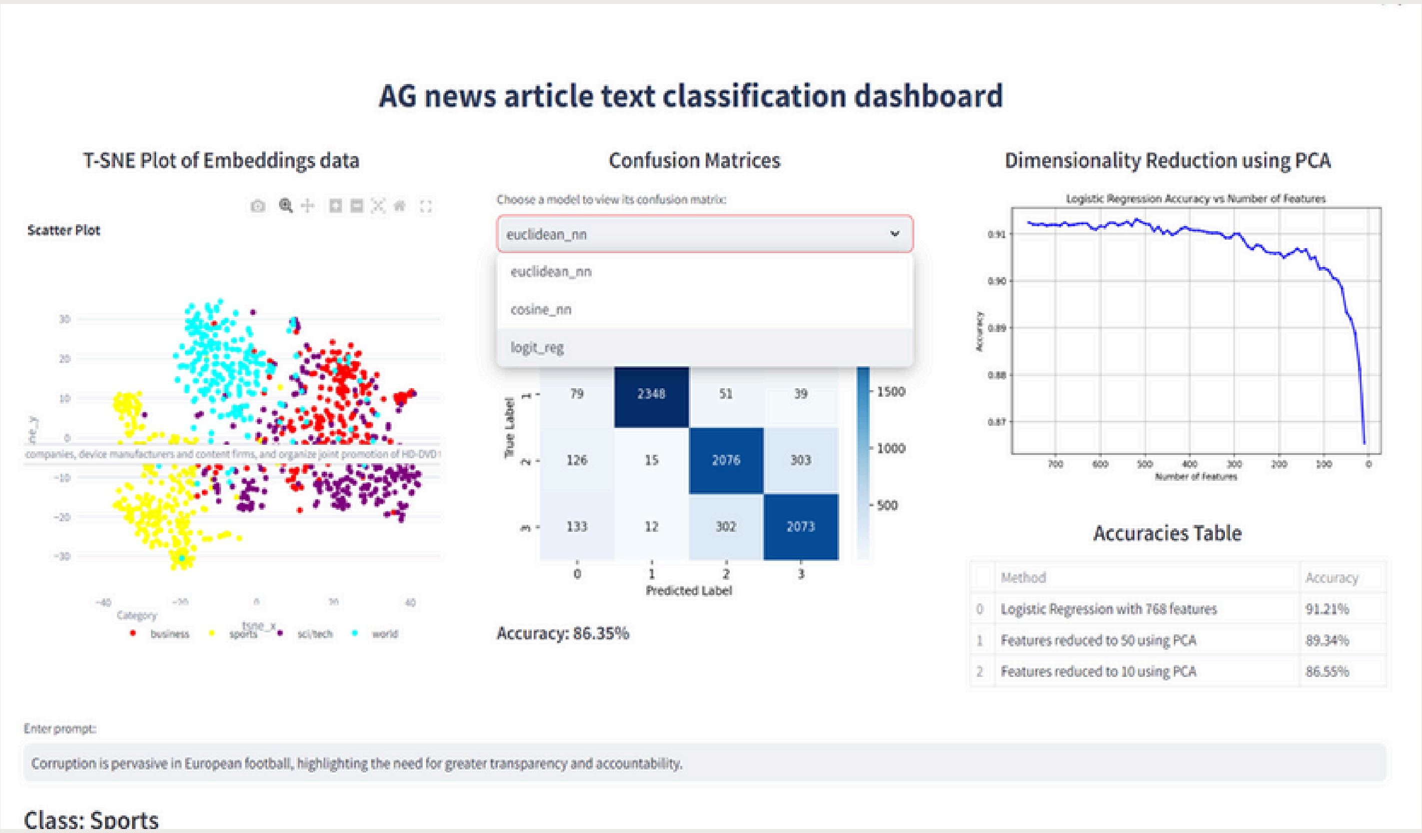
- Data Preparation: Loaded and shuffled the AG News dataset to ensure randomness. Created SQLite tables (raw_data and embeddings_data) to store text, labels, and embeddings.
- Text Embeddings and Processing: Used the DistilBERT model to convert text into 768-dimensional embeddings, processing data in batches of 512 with GPU acceleration for efficiency. Stored the embeddings in the database.
- Dimensionality Reduction and Visualization: Applied T-SNE to reduce embeddings to 2D for visualizing clusters and examining the separability of labels using scatter plots.
- Classification and Evaluation: Computed centroids for each class and classified test embeddings using Euclidean and Cosine distances. Compared the results using confusion matrices and accuracy scores.
- Outlier Detection: Identified outliers based on their distance from the class centroids.

- Logistic Regression and PCA: Trained a logistic regression model on the embeddings and compared its performance with nearest neighbor classifiers using accuracy and confusion matrices. Reduced the dimensionality of embeddings to 10 and 50 using PCA and evaluated the model's performance on the reduced data, balancing accuracy and computational efficiency.
- Insights from Dimensionality Reduction: Analyzed clustering patterns from t-SNE and PCA, studying their effect on classification accuracy.

Insights and Discoveries from Exploratory Data Analysis:

- Text Embeddings: Text data was converted into 768-dimensional embeddings using DistilBERT, allowing the model to analyze and detect patterns in the text numerically.
- t-SNE Visualization: t-SNE reduced the embeddings to 2D, showing clear clusters of labels, which demonstrated the model's ability to distinguish between classes effectively.
- Nearest Neighbor Classification: By calculating centroids for each class and classifying test data using Euclidean and Cosine distances, both methods achieved similar accuracy (86.35% and 86.34%), indicating consistent performance across distance metrics.
- Outlier Detection: Outliers were identified based on their distance from centroids, highlighting instances of misclassification or significant deviation from the clusters.
- Logistic Regression Classifier: Training on all 768 features resulted in a prediction accuracy of 91.2%, demonstrating the embeddings' strong representational ability for classification tasks.
- Impact of Dimensionality Reduction (PCA): Reducing the dimensions to 10 components maintained an accuracy of 86.55%, and with 50 components, the accuracy improved to 89.34%, showing that even with fewer dimensions, the embeddings still provide valuable predictive information.
- Accuracy vs. Features Trend: Adding more PCA features gradually improved accuracy, with diminishing returns beyond 50 dimensions, highlighting the balance between computational efficiency and model performance.

Interactive Dashboard for Data Exploration and Model Insights:



AG news article text classification dashboard

T-SNE Plot of Embeddings data

Confusion Matrices

Choose a model to view its confusion matrix:

euclidean_nn

True Label	0	1	2	3
1	79	2348	51	39
2	126	15	2076	303
3	133	12	302	2073

Accuracy: 86.35%

Dimensionality Reduction using PCA

Accuracies Table

	Method	Accuracy
0	Logistic Regression with 768 features	91.21%
1	Features reduced to 50 using PCA	89.34%
2	Features reduced to 10 using PCA	86.55%

Enter prompt:

Corruption is pervasive in European football, highlighting the need for greater transparency and accountability.

Class: Sports

The dashboard offers an interactive and visual way to explore the AG News dataset, focusing on text classification models and their performance.

- T-SNE Plot: A 2D representation of data points after dimensionality reduction, color-coded by category (World, Sports, Sci/Tech, Business). Users can interact by selecting/deselecting labels to explore clusters and hover over points to view the corresponding text.
- Model Selection and Confusion Matrix: A drop-down menu allows users to choose models (Euclidean Nearest Neighbor, Cosine Similarity, Logistic Regression), displaying the corresponding confusion matrix and prediction accuracy to assess model performance and errors.
- Dimensionality Reduction Accuracy Plot: This plot shows how accuracy changes with the number of features in the Logistic Regression model (from 768 to 10). It highlights the trade-off between model complexity and performance.
- Accuracy Comparison Table: A table summarizes prediction accuracies for different model configurations, helping users compare performance and understand the impact of dimensionality reduction.
- Prompt Bar: A prompt bar at the bottom allows users to input text, with the trained model providing predictions based on the selected category.

Overcoming Challenges: Solutions and Limitations in the Analysis:

- Understanding DistilBERT Embeddings: DistilBERT, primarily a text classifier, initially posed challenges in extracting meaningful embeddings for analysis. While its last layer outputs classification logits, these do not capture the text's semantic meaning. By using the second-to-last layer, which provides 768-dimensional embeddings, we were able to extract compact representations of the text for clustering and classification tasks.
- Tokenization Process: Tokenization is crucial in text preprocessing for NLP models. Incorrect tokenization can affect the final embeddings. To ensure accuracy, we used DistilBERT's pre-trained tokenizer, which automatically handles tokenization issues like padding and truncation, allowing us to focus on model training and evaluation.
- Vector Space and Relationships: Visualizing and interpreting the relationships between text embeddings is complex due to the nature of language. To address this, we used t-SNE for dimensionality reduction, visualizing embeddings in 2D to better understand how articles from different categories were grouped, highlighting the model's ability to distinguish between them.
- Language Complexity: Natural language’s inherent complexity and ambiguity posed challenges in accurate text classification. DistilBERT, though powerful, can struggle with words having multiple meanings or domain-specific terms. To address this, we utilized a pre-trained, fine-tuned model capable of handling general language understanding tasks.
- Limitations of the Analysis: The DistilBERT model, while effective, may struggle to generalize with domain-specific data, such as medical or legal texts, and could benefit from fine-tuning for such domains. Dimensionality reduction using PCA, while improving efficiency, can lead to a loss of information and reduce model performance as features decrease. The computational cost of working with large models like DistilBERT and techniques such as t-SNE for dimensionality reduction can limit scalability, especially with larger datasets. Additionally, while DistilBERT produces high-quality embeddings and predictions, its "black-box" nature makes it challenging to interpret, and techniques like SHAP values or LIME could help improve interpretability.

Conclusion:

This project successfully applied DistilBERT for text classification using the AG News dataset, leveraging text embeddings to effectively categorize news articles into distinct groups. Through techniques like t-SNE for dimensionality reduction and clustering, we visualized the relationships between different categories, providing valuable insights into the model's ability to distinguish between them. The interactive dashboard further enhanced the exploration of data, offering an engaging way to evaluate model performance and analyze clustering patterns. The analysis demonstrates the power of embedding-based models in text classification, providing a solid foundation for further development and potential real-world applications in news categorization and recommendation systems.