Automated News Classification Using Machine Learning

Member 1	I-An CHO	cianvt@vt.edu
Member 2	Richard LEE	richardva@vt.edu

Introduction and Project Problem Statement

Nowadays, many news aggregation platforms need an automated way to classify the large and growing amount of news. Although the original news resource might already contain category labels, the labels are not always consistent from various platforms. Therefore, an automated classification method is needed to enhance efficiency.

With the rapid update of news, the automatic classification system can classify news in real time to ensure that users get the latest and most relevant content. It also improves the resilience of news resources.

Data Set and Preprocessing Steps

Dataset: MIND: Microsoft News Dataset: https://msnews.github.io/

MIND has around 160k English news articles. Each article has details like News ID, Category, sub-category, title, abstract, description, and URL. Even if the total features aren't achieving 200, we enhance the dataset by doing feature engineering on the news content. Using Transformer-based models, such as BERT, we can extract contextual and semantic details from the text. These models give us embeddings with 768 dimensions, which show relationships between words in a sentence and the overall context. After getting 768 dimensions, PCA(principal component analysis) reduction will be utilized to decrease the dimensions. The final number of dimensions will be decided based on the best performance.

There are several steps to clean and pre-process the data. Firstly, to delete the row with missing values, the dropna() function will be utilized. Furthermore, news text will be encoded through text tokenization and obtained vector presentation by using a sequence classification model. Finally, PCA will be used to reduce the dimensionality of the 768-dimensional vector.

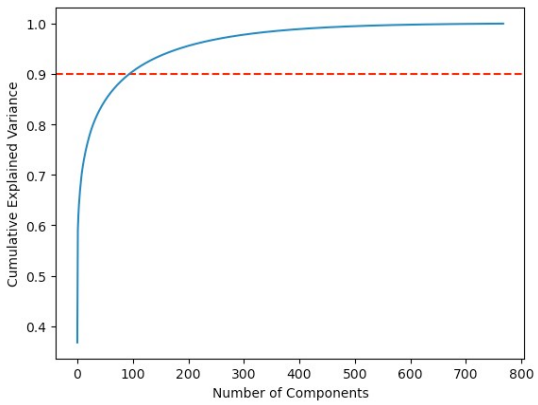
Model Input and Output

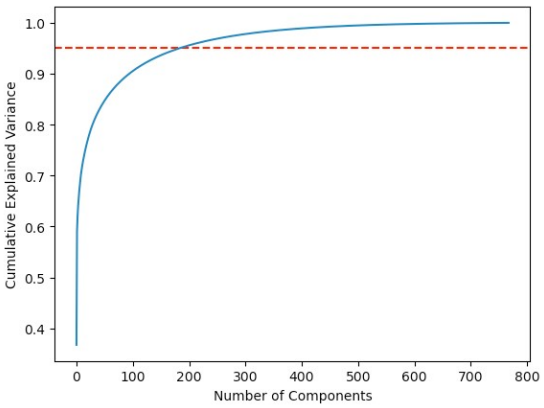
Input: abstract*, category(target)
*The 'abstract' of news articles as the main input, undergoing preprocessing to extract key features. This includes the use of transformer models (such as BERT) to obtain a 768-dimensional embedding representation of the text, subsequently reduced in dimensionality through Principal Component Analysis (PCA).

Output: The output will depend on the models, but it will include either of the following: The predicted news category or label; Possible category probability scores; If applicable, it might also provide additional information related to the classification.

PCA

To decide how many components we will use in the future, the Cumulative Explained Variance Plot was used. It can provide a visualization of the cumulative percentage of variance explained by each successive principal component. We initialized and fit a Principal Component Analysis (PCA) model with all (768) components to the embedded text data ('Abstract'). After fitting the PCA, we extracted the explained variance ratio for each component, which indicates the proportion of the total variance in the data that is captured by each principal component.





a. **Cumulative Explained Variance by Number of Components for 90% Threshold**
b. **Cumulative Explained Variance by Number of Components for 95% Threshold**

The number of components needed to explain **90%** of the variance is **95**
The number of components needed to explain **95%** of the variance is **184**

After determining the optimal number of components using the Cumulative Explained Variance plot for both 90% and 95% thresholds, we performed PCA on our data for each threshold using the respective number of components. Following this, we individually integrated the results of the two reduced-dimensional datasets back into our news data (2 copies) and saved them. With this preprocessing step completed, the data is now prepared and ready to be applied to a machine learning model.

Methods and Models

This is a classical supervised machine learning problem. The goal is to predict the categories of news articles. We use the labeled data to train the model and utilize this model to predict the new, unlabeled categories.

The models we are using:

1. Logistic Regression

2. Naive Bayes

3. K Nearest Neighbors

4. Decision Trees

Results

For PCA:

Cumulative Explained Variance Plot

For trained model test result:

Cross Validation, Confusion Matrices, ROC Curve, and AUC

Conclusion

What are the takeaways? How will your solution be impactful?

With Transformer-based models, we can extract contextual and semantic details from the text. Using PCA can reduce the features and identify underlying structures to make the model more precise. Finally, compare results from different machine learning models to find out the most optimal one. Our nuanced news classification model will be consistent across platforms, feature engineering powered, and versatile.

Individual Contribution (for groups only)

Member 1 I-An CHO	● Dataset researching ● Feature Engineering (PCA) ● Logistic Regression, Naive Bayes ● Documenting
Member 2 Richard LEE	● Methods and models researching ● Feature Engineering (Cumulative Explained Variance) ● K Nearest Neighbors, Decision Trees ● Result data visualization