News Classification System

Group 2

A0274955M A0286188L A0274722A

Dataset - https://huggingface.co/datasets/ag_news

Abstract:

With the enormous outburst in news content generated, it is important to classify the document/article into proper labels for brevity and ease of access as well as retrieval, thus organizing the content. This project aims to investigate multiple machine learning techniques to automatically classify the document into one of four distinct news classes: World, Sports, Business, and Science/Technology.

Goals:

1. Pre-processing of Text
2. Coming up with efficient and apt representations/embeddings of the words/passages in the text
3. Developing a model for the classification of news

Brief Summary:

1. Pre-processing: The provided text has special characters, stopwords, and special tags like backslashes(“/”), “<>” brackets, and many more. Therefore, it is necessary to remove such unwanted information for more accurate representations. This also includes the need to balance the dataset to tackle class imbalance and biased learning
2. Representations/Embeddings - To have a relation between semantic words of similar meaning, it is necessary to have word embeddings. This will also help in out-of-vocabulary handling. Doc2Vec and Word2Vec models will be explored in this regard.
3. ML/NLP Models - As text is sequential data, various models can be experimented with and further fine-tuned. The Long Short-Term Memory model works well on textual data. Along with that, we will be developing an RNN model and comparing various metrics including accuracy, f1 scores, etc of these developed models to BERT Transformers and provide an analysis.

Potential Problems:

1. Semantic Complexity of Sentences - Often text is not as simple as it seems and people often embed some hidden semantic meaning which requires a deeper level of Grammatical Understanding. Tackling this will surely be difficult and will play a major role in deciding the embeddings to be followed.
2. Domain-Specific Vocabulary- We cannot be sure that the new news article for prediction only consists of terms the model has learned. Tackling new terms like new Cooperations, Player names, etc will pose a challenge.
3. Handling Outliers - There might be news articles having an overlap of terminologies of two or more domains. Handling such cases will pose a challenge.