**TWITTER FINANCIAL NEWS CLASSIFICATION**

**PROJECT OVERVIEW**

**Title:** Twitter Financial News Classification Using Machine Learning and Deep Learning

**Colab Link:** [**https://colab.research.google.com/drive/1DzvefJx86txYe9NeFouyGSfEwchpoI9S?usp=sharing**](https://colab.research.google.com/drive/1DzvefJx86txYe9NeFouyGSfEwchpoI9S?usp=sharing) **GitHub Link:** [**https://github.com/PrernaMaurya/Twitter-Financial-News**](https://github.com/PrernaMaurya/Twitter-Financial-News)

**Objective:** To develop a model that accurately classifies finance-related tweets into 20 predefined categories, aiding in the organization and analysis of financial information disseminated on Twitter.

**Dataset Summary:**

**●Source:** Finance-related tweets

**●Total Tweets:** 21,107

**●Training Set:** 16,990 tweets

**●Validation Set:** 4,117 tweets  
**●Categories:** 20 financial topics including Stock Commentary, Earnings, IPO, M&A, Macro, Gold, Crypto, Energy, etc.

DATA PREPROCESSING  
**Steps Undertaken:  
●Text Cleaning:** Removed URLs, punctuation, numbers, hashtags, mentions, and converted text to lowercase.

**●Tokenization & Padding:** Utilized Keras' Tokenizer and pad\_sequences to prepare text data for deep learning models.

**●Label Encoding:** Applied one-hot encoding to convert categorical labels into a binary matrix.

**Observations:** The preprocessing steps effectively prepared the textual data for model training, ensuring consistency and facilitating better learning.

**EXPLORATORY DATA ANALYSIS (EDA)**

**Techniques Used:**

**●Label Distribution:** Visualized using bar plots to understand the frequency of each category.

**●Word Clouds:** Generated for each category to identify common terms and gain insights into the textual data.

**Insights:** The dataset exhibits a mostly balanced distribution across categories, which is favorable for training classification models.

**MODEL DEVELOPMENT**

**Models Implemented**

**1. Logistic Regression:**

**●Features:** TF-IDF vectors

**●Purpose:** Served as a baseline model for comparison.

**2. Artificial Neural Network (ANN):**

**●Architecture:** Input layer → Dense layers with ReLU activation → Output layer with Softmax activation.

**●Features:** TF-IDF vectors

**3. Convolutional Neural Network (CNN):**

**●Architecture:** Embedding layer → Convolutional layers → MaxPooling → Dense layers → Output layer.

**●Features:** Tokenized and padded sequences

**4. Recurrent Neural Network (RNN) with LSTM:**

**●Architecture:** Embedding layer → LSTM layers → Dense layers → Output layer.

**●Features:** Tokenized and padded sequences

**Training Details:**

**●Loss Function:** Categorical Crossentropy

**●Optimizer:** Adam

**●Metrics:** Accuracy

**●Epochs:** Varied per model (e.g., 5-10 epochs)

**●Batch Size:** Varied per model (e.g., 32-64)

**MODEL EVALUATION**

Each model was evaluated on the validation set using accuracy, confusion matrices, and classification reports. Here's a summary of validation accuracies:

|  |  |
| --- | --- |
| MODEL | VALIDATION ACCURACY |
| Convolutional Neural Network (CNN) | 82.3% |
| Artificial Neural Network (ANN) | 80.9% |
| Logistic Regression | 77.5% |
| Recurrent Neural Network (RNN) | 20.7% |

**Observations:**

●CNN and ANN achieved high accuracy, showing strong performance on short-text classification.

●The RNN significantly underperformed, likely due to its shallow architecture, limited training epochs, or the fact that CNNs handle short text patterns more efficiently.

●Confusion matrices revealed most misclassifications occurred between semantically similar financial categories.

**MODEL COMPARISON**

**Visualization:**

●Created bar plots to compare validation accuracies across all models.

**Insights:**

●Deep learning models (CNN, ANN) outperformed the traditional Logistic Regression.

●RNN performance was significantly lower, reinforcing that CNN is better suited for short-sequence text like tweets.

**CONCLUSION**

This project successfully implemented and compared multiple models for classifying finance-related tweets into 20 categories. The deep learning models outperformed traditional logistic regression, with CNN achieving the highest validation accuracy (82.3%). This demonstrates the strength of CNNs in capturing local patterns in short financial texts.

Although RNN underperformed, its inclusion in the study highlighted the importance of model architecture and training depth in NLP tasks. The project showcases the effective use of Natural Language Processing in organizing real-world financial data and sets the stage for further improvement through hyperparameter tuning or advanced models like BERT.

**FUTURE WORK**

**●Model Optimization:** Experiment with hyperparameter tuning, dropout rates, and deeper layers.

**●Advanced Architectures:** Implement transformer-based models like BERT for even better performance.

**●Data Augmentation:** Use more diverse and updated financial tweets for broader coverage.

**●Deployment:** Package the model into an interactive web app using Streamlit (optional).

**●Explainability:** Add model interpretability tools like LIME or SHAP to explain predictions.

**NAME- PRERNA MAURYA  
EMAIL:** [**prernam0203@gmail.com**](mailto:prernam0203@gmail.com) **DATE OF JOINING: 15th April 2025  
DURATION OF INTERNSHIP: 3 Months**