# **HW Assignment 2 (A2): Deep Learning**

# CS6120: Natural Language Processing Spring 2024

Due Date: Tuesday,

**(10 points)**

# One price for unlimited news trade categories - Key FeatureHuffPost Shuts Down Unpaid Contributor Blogger Program

# Scenario

# HuffPost (known as The Huffington Post until 2017) is a prominent website that provides a mix of news, commentary, and various forms of original content, including blogs and articles covering a wide range of topics from politics and business to entertainment and lifestyle.

# The website initially featured contributions from unpaid bloggers from diverse fields such as politics, entertainment, and academia, amassing around 100,000 contributors by 2018. Over the years, the site has seen involvement from notable figures, including celebrities, politicians, and academics, who have contributed content, enhancing its reputation and reach. HuffPost has undergone significant transformations since its inception, expanding its content to include more news-focused articles and reducing reliance on its unpaid blogger program. It now features commissioned opinion pieces and personal essays. The website operates on a revenue model derived from advertising and maintains its content as free to access for users. Some fascinating history on mergers and acquisitions: The Huffington Post was acquired by AOL in March 2011 for $315 million, which expanded HuffPost's reach and resources. In 2015, Verizon Communications acquired AOL, and HuffPost became part of Verizon Media. HuffPost rebranded in April 2017, including updates to its website design, logo, and content approach. In February 2021, BuzzFeed acquired HuffPost from Verizon Media in a stock deal, marking another significant shift in the platform's ownership and business strategy.

You are assigned a classification task through a large dataset from HuffPost. Classification tasks lay the foundation for many applications that involve understanding, interpreting, and generating human language.

Some prominent examples of classification techniques include:

* Sentiment Analysis
* Spam Detection
* Topic Categorization
* Language Identification (used to determine the language of a given text automatically)
* Named Entity Recognition: (classification plays a role in determining the category of identified entities such as person, organization, location, etc. within text)
* Intent Classification (In conversational AI and chatbot development classifying the intent of a user's input allows the system to provide appropriate responses or actions)
* Fake News Detection
* Authorship Attribution (By classifying texts based on writing style, vocabulary use, and other linguistic features, algorithms can attribute anonymous or disputed texts to specific authors)

# Tasks

You are given more than 200k news items from HuffPost. The dataset is taken from a Kaggle competition:

The news articles in the dataset are each tagged with one of several categories (such as Politics, Technology, Entertainment, etc.). Each entry in the dataset typically includes the headline and a short description of the article, along with its assigned category/ The primary objective is to develop and compare various machine learning models to categorize news articles into predefined categories based on their headlines and short descriptions. This assignment will help you understand text classification nuances using traditional machine learning and deep learning approaches.

**1. Data Preprocessing:**

* Load the dataset and perform initial exploration to understand its structure.
* Clean the text data, including removing special characters, stopwords, and applying lowercasing.
* Perform text tokenization and vectorization using techniques like TF-IDF.
* Extract and analyze different features from the text that might be useful for classification, such as word count, sentence length, n-grams, etc.

**2. Model Implementation and Evaluation:**

* Divide the dataset into training and testing sets.
* Implement the following models and evaluate their performance on the test set using metrics such as accuracy, precision, recall, and F1-score:
* Logistic Regression: Use as a baseline model to understand the linear separability of text categories.
* Random Forest (RF): Explore ensemble methods in handling high-dimensional text data.
* XGBoost:
* Artificial Neural Network (ANN): Design a simple feedforward neural network with at least one hidden layer to classify news categories.
* Convolutional Neural Network (CNN): Implement a CNN to capture local dependencies in text data. Discuss the choice of the architecture in the context of text data.

**Optionally**

Long Short-Term Memory (LSTM): Use LSTM to model dependencies in text sequences.

**3. Comparative Analysis:**

* Compare the performance of conventional ML techniques with deep learning techniques. Discuss the trade-offs involved, such as computational complexity, model interpretability, and performance.
* Visualize the results using confusion matrices (and optionally ROC curves) for each model.

**4. Discussion:**

* Discuss the challenges encountered during the implementation of each model, including issues related to overfitting, underfitting, and model tuning.
* Reflect on the importance of preprocessing and feature engineering in text classification tasks.

# **Expected Output**

Please submit a fully executed jupyter notebook identifying question number and steps. Make sure to add comments to your solution.