Python/ Deep Learning Project Proposal

By

Team No: 2

Member 1: Akhil Teja Kanugolu Class ID: 11

Member 2: Geetanjali Makineni Class ID: 13

TITLE:

To identify the type of the news based on the headlines and short descriptions.

MOTIVATION:

The most crucial source that plays an important role these days for knowing what is happening around is NEWS. Nowadays, we have many sources on the Internet that are generating a high amount of news. Adding on to that, the importance for knowing the information by the users/people is rapidly increasing. So, it is important that the news being classified into categories to allow the users accessing the information that they are more interested in. This helps them to jump into the category that interests them. Thus, the machine learning model for the automated classification of news can be used for identifying the topics for untracked news and make the users suggestions depending on their prior interests.

OBJECTIVE:

The main objective of our project is to take the headlines as well as short descriptions from the news articles as the input and give the output based on the category like (politics, entertainment, health, travel, etc.,) to which the news belongs to. Here, we use different types of machine learning algorithms that helps us finding which model gives better performance. We also use Convolution Neural Network and find which model can yield better performance here by measuring accuracy.

DATA SET DESCRIPTION:

The dataset which we use here is taken from Kaggle which contains around 200k headlines of news from the years 2012 till 2018. These are obtained from the webpage Huffpost. News in the dataset belong to around 40 different labels. Each one of the records consists several types of attributes. But, from all the attributes we just consider taking only the Headline, Category and Short description for our further analysis. Here, we also combine two data attributes namely Headline and Short Description into one single attribute for using as an input.

FEATURES:

The whole idea is divided into 3 main components as below:

- Dataset Preparation – Here, we process by loading a dataset and perform pre-processing by replacing them with median or mean and then split the data into train and test validation sets.
- Feature Engineering – In this step, the raw dataset is transformed further into flat features. This mainly includes process in which we create new features from the existing features. We use the Count Vector matrix notation where we check variance and drop some of the features based on the threshold it handles.
- Model Training – The final most step we use is the Model Building where a machine learning model is here trained on the dataset. We implement the models like Naïve Bayes Classifier, Convolution Neural Network and Decision Tree Model and find which model gives best accuracy.

SOFTWARE AND PLATFORMS:

- Jupyter Notebook
- Google Colab

REFERENCES:

- https://www.kaggle.com/rmisra/news-category-dataset
- https://www.huffpost.com/