ABSTRACT

NetApp Data Challenge

DATA WARRIORS

(Team-ID: TM190904)

Team Members:

Bhargav D

Prerit Jain

Shubham Mawa

PROBLEM DESCRIPTION:

The problem is to analyze the text data i.e News headlines and the metadata i.e the short description to classify the news document into pre-defined fixed category set which showcases the broad topic discussed in the news document. The model would be expected to guess the correct category for the correct news headline category of data not necessarily belonging to the same source of the News Headlines Dataset.

DATASET EXPLORATION:

The training set contains 202,372 records. Each JSON record contains the following attributes :

- id: ID of the News Headline.
- category: Category of the News Headline.
- headline: The News headline.
- short_description: Short description of the News Article

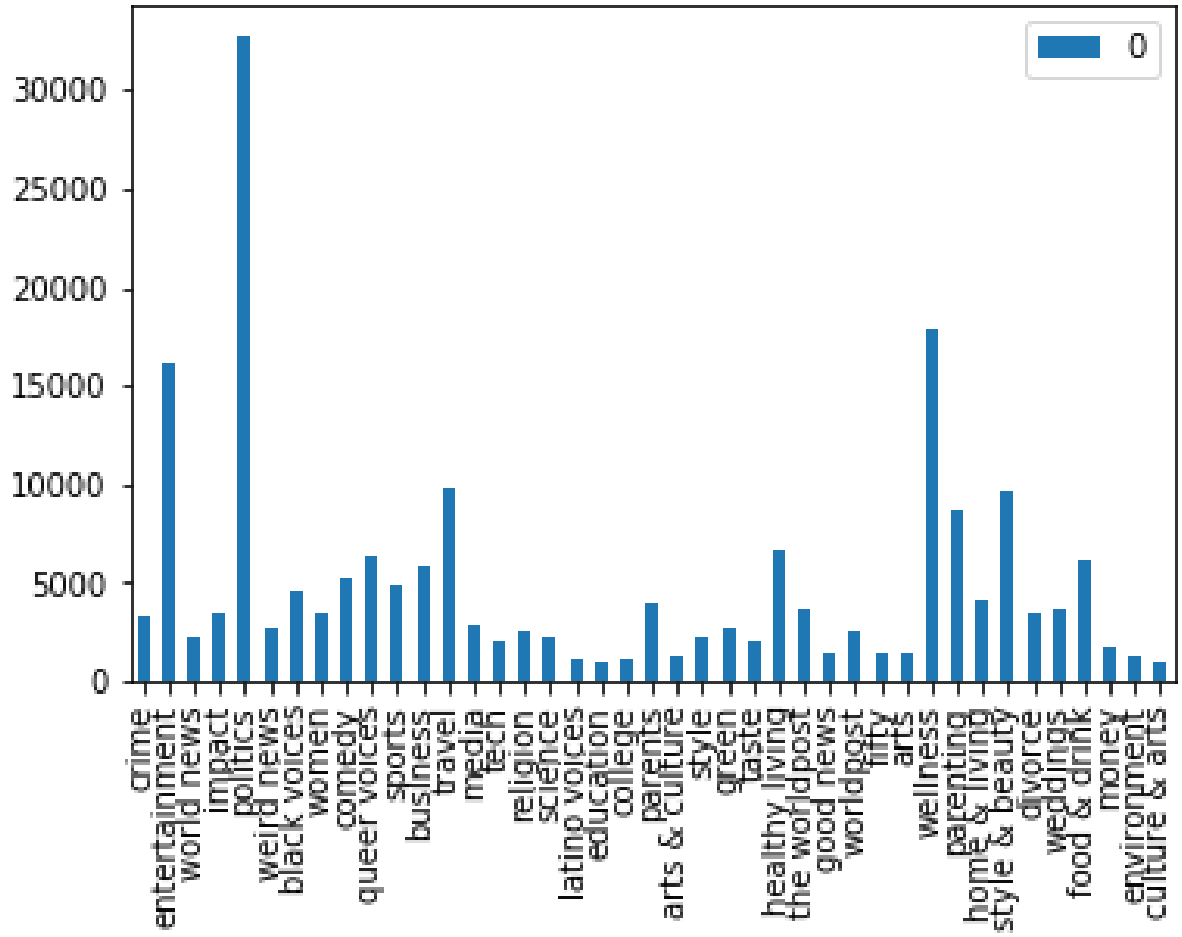
The Model we learn should be able to predict the category of a news article given its headline and it's corresponding short description.
There are 41 unique output category variables.

The top 5 categories of the News Articles in terms of frequency are:

1. Politics
2. Wellness
3. Entertainment
4. Style and Beauty
5. Travel

Some of the low frequency occurring categories were culture & arts, education, college, good news, environment.

These results reinforce what we already know, News articles nowadays only care about the clicks they get or how attractive the headline is rather than being about topics which influence the society in a positive manner.



Bar Graph of Frequency of News Articles versus Output Category of the News Articles

METHODOLOGY:

The task is to predict the categories of the News headlines which generates the need for representing the headlines into dense paragraph embeddings for better analysis.

- Word embeddings will be generated using the fasttext toolkit which generates dense vectors of dimension 300 using Skip-gram model or DBOW(Distributed bag of Words).
- The embeddings will be fed into a neural network architecture which will predict the category of the given News Headline. Our neural network model will be a combination of **CNN(Convolutional Neural Network) and Bi-LSTM (Bidirectional Long Short-Term Memory)**. We expect that the combination of CNN and Bi-LSTM will capture different information which will help to enhance the overall performance in the given multi-class classification task.
- In one approach the word vectors will be kept static and in another, the word vectors will also be fine-tuned when the model is being trained. Retraining the word vectors might help a little as they will become more task-specific but at the same time could affect how well the model generalizes to various other News documents. We plan to test both models and see which gives better performance on unseen data.
- The CNN creates feature maps which can capture trigger words which are sometimes important and are causing the headline to be classified as of a particular category. The Bi-LSTM network captures dependencies which are important for linking events in a news description and subsequently accurate category classification.
- The outputs of the CNN and Bi-LSTM will be combined and fed into a fully connected network and finally, a softmax layer which computes the multi-class probabilities and the prediction is done.

Traditional machine learning models are not able to capture such information as is possible with deep neural network models which capture low-level to high-level abstractions of the input data and complex features are learned in deeper layers which otherwise would have to be created manually if traditional machine learning models are used, which is a very tedious and inefficient process and will in most cases lead to poorer results. The training dataset is also quite large so the problem of a deep learning model overfitting the data is also not there and hence the use of a deep neural network model is highly encouraged.

NEURAL NETWORK ARCHITECTURE