AUGUST 31, 2024

https://github.com/hanhng294

SUMMARY OF FINDINGS

NEWS CLASSIFICATION

Written By : Grace Nguyen

Classifying news category based on title

The dataset News Aggregator provided title and category, which we can utilize to build machine learning models

The objective of this project is to build a predictive model that classifies news category based on title. With the vast amount of information, the automation of category classification for news articles could save time, effort and set forth for recommendation system.

We used NLP techniques and deep learning model (LSTM) to achieve this goal

Results

- Model: LSTM, stratify based on category to ensure equal distributions
- Result: we achieved roughly 0.95 for train set and 0.90 for validation set

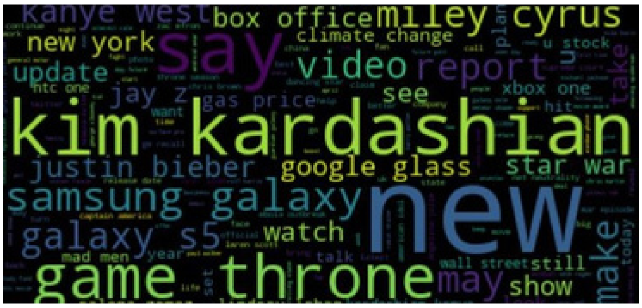
for more details: check out the .ipynb file

Exploratory Data Analysis

- Dataset information: there are 422,419 rows with 4 categories (business, technology, entertainment, health).
- Distribution: by using bar chat, we can see that entertainment (e) is the most popular category while the least is health (m)
- Popular words: among the 69,293 distinct words, some most common words are new, google, apple, etc.

Text Preprocessing

- Text preprocessing: lowercase, remove punctuation, remove stop word, lemmatize (to convert text to its original forms)
- Tokenize, sequencing and padding for predictors to convert text to integers in the form of equal-length arrays
- Label encode for target variables



Word cloud visualizing common words