News Category Classification

Group Name: G

Group Members:

First name	**Last Name**	**Student number**
Dev	Makwana	C0885064
Joel	Crasto	C0883863
Krina	Patel	C0886861
Mahaveersinh	Chauhan	C0884854
Trushna	Patel	C0886910

Submission date: 17/08/2023

Table of Contents

Abstract

Effective ways for organizing and categorizing enormous amounts of textual material are required in the current digital era due to the explosive increase of online news items. This project investigates Natural Language Processing (NLP) with an emphasis on news Categories Classification, with the objective of automatically classifying news articles into pertinent and significant categories. This study's main goal is to create and put into use a reliable classification model that can correctly classify news articles into predetermined groups, facilitating effective content organization and information retrieval. The fast expansion of internet news sources has fundamentally changed how we obtain and use information. The enormous number of news stories that are released every day calls for creative methods to efficiently classify and arrange these texts, enabling smooth access for users looking for information. We want to develop a complex system capable of automatically classifying news stories into a variety of categories, including politics, sports, technology, entertainment, and more by utilizing the strength of NLP approaches. This categorization not only makes it easier to find relevant material, but it also provides insightful information about current events, popular opinion, and theme patterns in news items.

Introduction

The demand for effective ways to organize and categorize enormous amounts of textual data has never been greater than it is now, in an age marked by an overwhelming influx of digital information. The difficulty of categorizing news stories into pertinent and meaningful categories has emerged as a crucial job at the nexus of technology and information transmission with the introduction of online news platforms. The goal of this project is to revolutionize how we process and consume news content in the digital era.

The study starts out by looking into different NLP methods, such as text preprocessing, feature extraction, and machine learning algorithms. Raw text is transformed into numerical representations using feature extraction techniques like TF-IDF (Term Frequency-Inverse Document Frequency), making it easier to train classification algorithms. Several machine learning algorithms, ranging from traditional approaches such as CatBoost and LightGBM to more advanced deep learning techniques like Recurrent Neural Networks (RNNs), are implemented and evaluated for their performance in news categorization.

Furthermore, the project delves into model optimization techniques, hyperparameter tuning, and ensembling strategies to enhance classification accuracy and robustness. The experimental results showcase the efficacy of the proposed classification model in accurately categorizing news articles across a range of domains, including Business, Politics, Food & Drink, Travel, Parenting, Style & Beauty, Wellness, World news, Sports, and Entertainment. The project demonstrates the practicality and relevance of NLP techniques in automating news categorization, ultimately streamlining the process of information retrieval and content organization in the digital news landscape. This project comprehensively studies News Categories Classification using state-of-the-art NLP methodologies. The findings contribute to the advancement of NLP research and hold significant implications for real-world applications, paving the way for more efficient and effective news content organization and retrieval systems.

Methodology

The goal of the project is to classify the categories of news articles based on the analysis of the description provided using Natural Language Processing. The basic steps implemented for creating the project are outlined below:

1. Dataset details

This dataset used in this assignment is from the Kaggle (*News Category Dataset*, n.d.) platform that consists of some short descriptions of the news articles. The dataset contains 50000 news headlines from the year 2012 to 2018 obtained from HuffPost. The categories are Business, Politics, Food & Drink, Travel, Parenting, Style & Beauty, Wellness, World news, Sports, and Entertainment. Table 1 describes features and their meaning for analyzing and classifying:

Field Name	Description
category	The category in which the article was published.
headline	The headline of the news article.
links	The link to the original news article.
short_description	Abstract of the news article.
keywords	The words extracted from the news links

2. Analysis Process

I. Data Cleaning

- **Dropping Null values:** The dataset from Kaggle (*News Category Dataset*, n.d.) contains 2668 null values that must be treated to get better results. Thus, they were removed as the dataset was large enough which will not affect much.
- **Remove Duplicate Values:** It is very crucial to check for duplicates and treat them to avoid ambiguity during training. Thus, 3962 duplicates were ignored which resulted in 43, 370 records remaining in the dataset.
- **Remove Irrelevant Columns:** The columns that seems to be irrelevant needs to be removed. In this case, the column links had the link of the news which does not impact the category so it was dropped from the dataset.

II. Text Preprocessing

The text was full of stop words and irrelevant words that did not impact the category. All these words need to be removed to create a clean text with the most relevant words. Thus, the text preprocessing process was implemented.

- **Stopwords Removal:** Stopwords are frequently used words that have little or no meaning (such as "the," "and," and "is"). Eliminating these terms helps the text read more clearly and concentrate on the words that have greater meaning.
- **Remove Punctuations and Special Symbols:** Special characters and punctuation such as, '@', '#', '$', ',', '.', and others can interfere with text analysis. The text can be made more typical by removing them. This step was mandatory to carry out as the text included apostrophes and so on.
- **Removing language other than English:** There are situations when different languages may be used in news descriptions. Language detection algorithms can assist in determining the language used in the text.

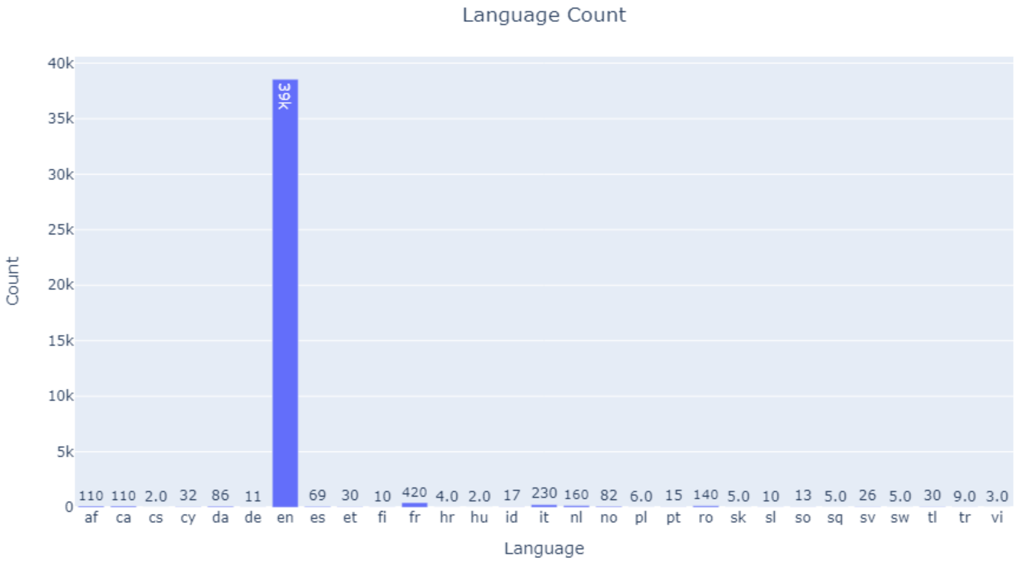

Figure 1 Language count of news

- **Removing miss-spelled words:** In order to ensure correct analysis, it is crucial to check the spelling of English terms. Spelling errors might produce inaccurate findings and interpretations. After removing special characters, some resulting words were misleading such as ‘ive’, ‘aaa’ etc.
- **Removing empty strings:** Some text may become empty strings as a result of the aforementioned actions. For example, strings in the headlines where the text was

initially of short length, applying above steps reduced the length to 0 which were removed.

- **Lemmatization:** Lemmatization is the process of breaking down words into their most basic or root form. This aids in standardizing terms like "resting" to "rest" and "better" to "good" that have similar meanings but distinct spellings.

III. Exploratory Data Analysis

1. Unigrams and Bigrams of news descriptions:

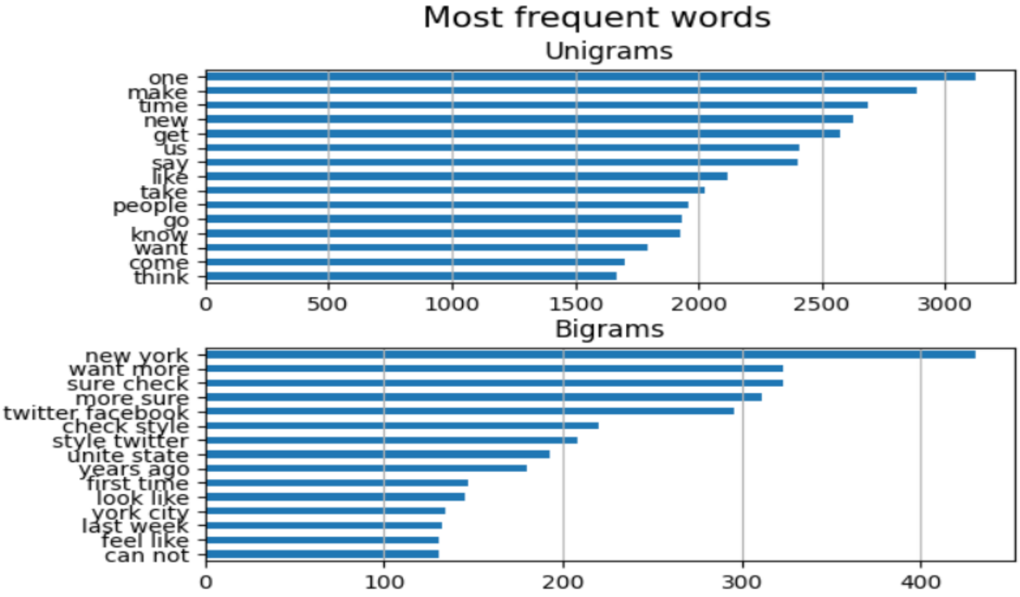

Figure 2 Unigrams and Bigrams of news description

2. Word Cloud of short_description feature:

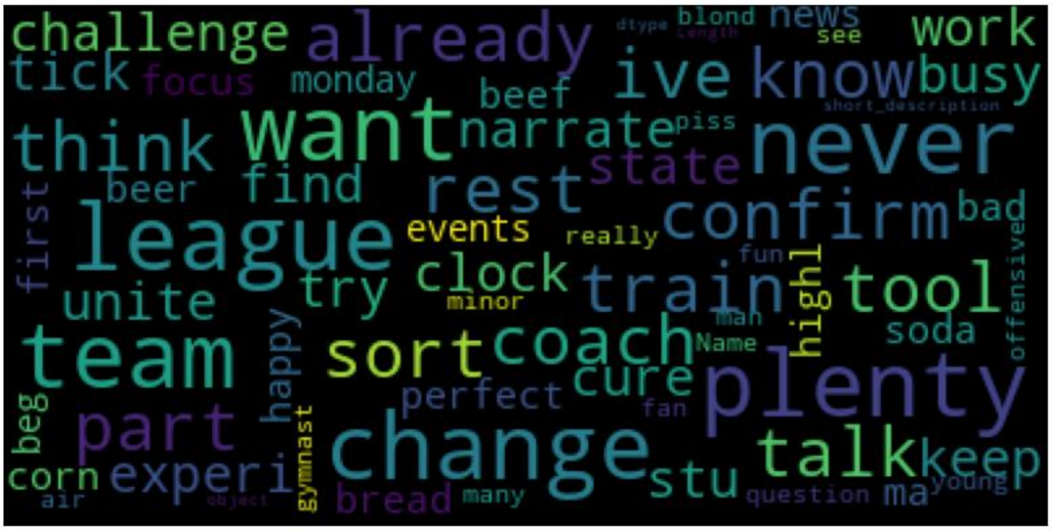

Figure 3 Word Cloud of short_description feature

3. Word Cloud of news headlines:

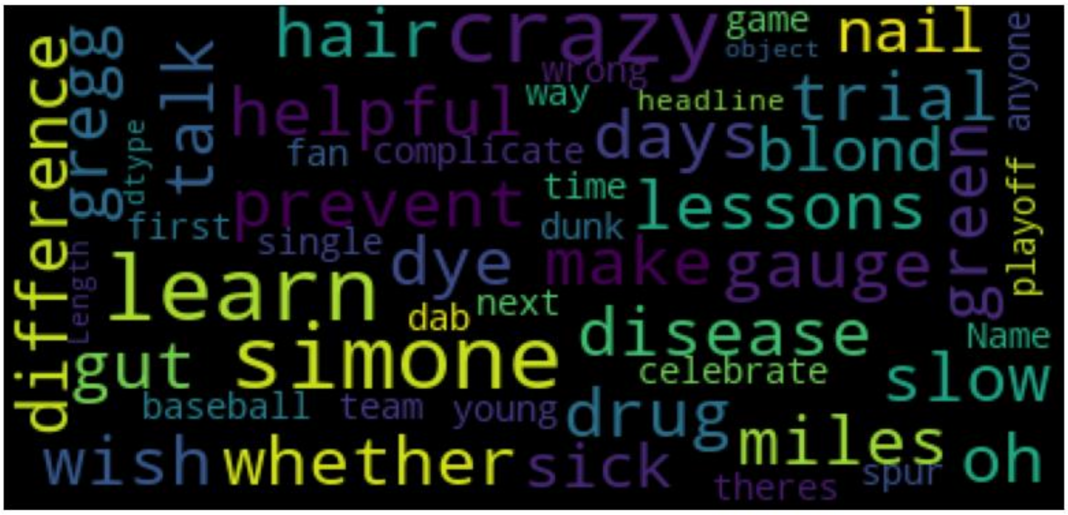

Figure 4 Word Cloud of news headlines

4. News Count by Categories

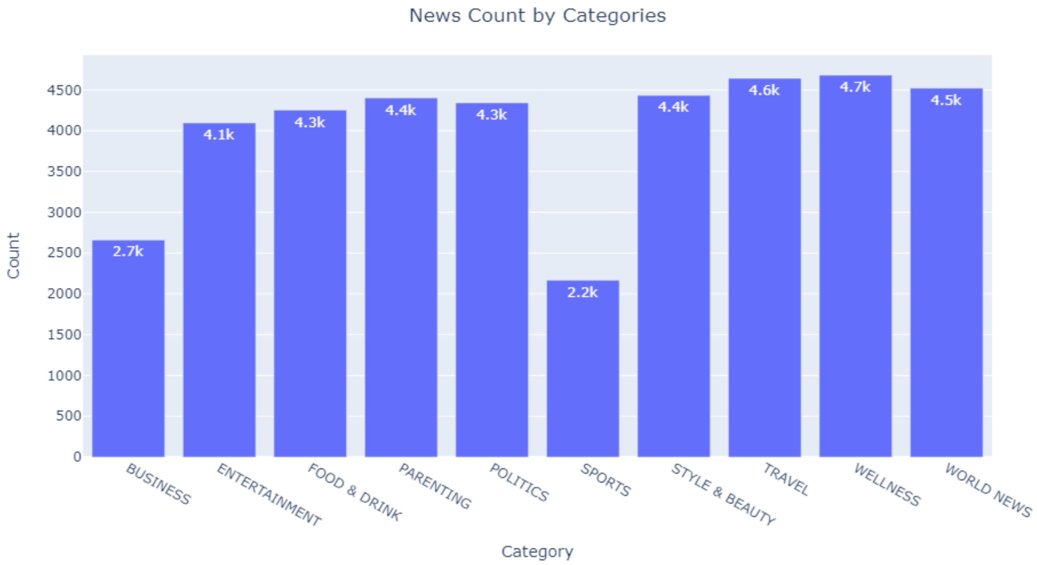

Figure 5 News count by category

IV. Feature Engineering

- **TF-IDF Vectorization:** Term Frequency-Inverse Document Frequency, or TF-IDF, is a method for transforming text input into a form that machine learning algorithms can understand. It illustrates how significant a word is in a document in relation to a group of papers. With values representing the significance of each word in each document, the outcome of this stage is a matrix with rows denoting documents and columns denoting distinct words. These operations resulted in approximately 20, 000 features from which, only 500 max_features were passed to PCA according to the TF-IDF scores.

- **Incremental PCA:** Traditional PCA is computationally expensive and requires more memory. In order to overcome these difficulties, a PCA variant called incremental PCA (IPCA) processes data in pieces or batches. When the complete dataset cannot be stored in memory at once or when processing speed is an issue, it is especially helpful. IPCA works in news category classification by dividing news descriptions into smaller batches, initializing, updating principal components, and dimension reduction. After applying IPCA, the dimensions of dataset reduced to 61 columns that contains the data of both headlines and descriptions.

V. Classification Models

- **Light Gradient Boosting Machine (LightGBM):** Text classification on news articles, is a well-known gradient boosting framework to be very effective and optimized for huge datasets. In terms of speed and memory effectiveness, LightGBM excels. It is ideally suited for text data with categorical features like word embeddings or word frequencies since it can handle categorical features directly without needing one-hot encoding.

- **Categorical Boosting (CatBoost):** It is a potent gradient-boosting algorithm that emphasizes its prowess in handling categorical features (Revert, 2020). Without the requirement for manual preprocessing, such as one-hot encoding, CatBoost is specifically made to handle categorical features. To handle categorical data in a straightforward manner, it makes use of a cutting-edge technique termed "ordered boosting." CatBoost uses methods like ordered boosting, which automatically manages the model's complexity and reduces the likelihood of overfitting. This can be quite helpful when working with text data with a large dimensionality. Through the class_weights parameter, CatBoost enables the correction of class imbalance, which may enhance the classification of underrepresented categories.

- **Extreme Gradient Boosting (XGBoost):** It is a common gradient boosting method (Verma, 2022) that excels at a variety of classification-related machine learning problems and has gained popularity for both. It assembles a collection of weak learners, typically decision trees, to produce a robust predictive model. In addition to handling categorical features utilizing methods like one-hot and integer encoding, XGBoost offers alternatives for handling numerical features. Decision trees are generated by XGBoost, and you can view individual trees to better understand how the model makes decisions. Understanding how news articles are categorized can be facilitated by this in particular.

- **Bidirectional Long Short-Term Memory (LSTM):** LST (How to Code RNN and LSTM Neural Networks in Python, n.d.) is a particular form of recurrent neural network (RNN) architecture that is well suited for text categorization tasks like identifying different categories of news. LSTMs can handle the sequence structure of time series data better than traditional feedforward neural networks since they have feedback connections. They can spot long-term connections and patterns, making them ideal for forecasting time series data. Bidirectional LSTMs process forward and backward sequences, allowing the model to capture contextual information from past and future words.

Results

From the above results, CatBoost Classifier provides better accuracy amongst all of the three algorithms. The CatBoost and LightGBM techniques were used to aim for a precise classification of classes as they both can work on categorical features. However, CatBoost can be seen to have some good results than others with the accuracy of 0.81 and 0.56 for training and testing, respectively. The model following it that gave better results was LightGBM with 0.87 and 0.55 for training and testing, respectively.

Baseline Model	**Accuracy (Training)**	**Accuracy (Testing)**
CatBoost	0.81	0.56
XGBoost	0.95	0.54
LightGBM	0.87	0.55

Thus, it was decided to perform hyper-parameter tuning for the CatBoost and LightGBM to set parameters in such a way that better and more efficient results can be seen. In the case of hyper-parameter tuning, the technique called GridSeacrhCV is used for its quick response. The parameters to tune for both models were not altered much as they require more resources to avoid Memory Error. In contrast to base models, the CatBoost model moves more towards overfitting while the LightGBM is providing a better fit results of 0.63 and 0.53 for training and testing.

Hyper Parameter Tuning	**Accuracy (Training)**	**Accuracy (Testing)**
CatBoost	0.88	0.56
LightGBM	0.63	53

In order to improve the classification results, the next step was to try a Bi-directional LSTM as it preserves the context of the words preceding and succeeding the current word. Hence, the Bi-directional LSTM was trained with 10 epochs and a learning rate of 0.01 with an Embedding layer that takes input as word embedding. The results were much improved in the case of Neural Networks. The final accuracy of Bi-directional LSTM was observed to be 0.63

Conclusion and Future Work

In conclusion, the process of categorising news using three well-known gradient boosting algorithms like, CatBoost, LightGBM, and XGBoost along with a combination of data and text preprocessing steps has shed important light on the efficiency of various approaches to this task. The unstructured news descriptions were converted into a structured format appropriate for analysis through thorough data cleaning, which included removing stopwords, and special characters, and completing lemmatization. Our findings show that each of the utilised algorithms has advantages and disadvantages when it comes to categorising news. In terms of training time and predicted accuracy, CatBoost, XGBoost, and Light GBM demonstrated outstanding efficiency, demonstrating its potential for real-time or resource-constrained applications. The LSTM model, a deep learning architecture created especially for sequential data, on the other hand, had an outstanding performance in capturing the complex temporal dependencies present in news stories.

Although our project has provided insightful information, there are still a number of opportunities for further investigation and improvement. Look at how ensemble approaches, like stacking or blending, can be used to combine the best features of various models and improve overall predictive performance. Utilise contextualized embeddings and incorporate cutting-edge pre-trained language models, such as BERT, GPT-3, or their successors, to improve classification accuracy.

References

News Category Dataset. (n.d.). News Category Dataset | Kaggle. https:///datasets/setseries/news-category-dataset

What is XGBoost? (n.d.). NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/data-science/xgboost/

What Is CatBoost? (n.d.). Built In. https://builtin.com/machine-learning/catboost

Revert, F. (2020, May 22). *Why you should learn CatBoost now*. Medium. https://towardsdatascience.com/why-you-should-learn-catboost-now-390fb3895f76

Verma, N. (2022, September 7). XGBoost Algorithm Explained in Less Than 5 Minutes. Medium. https://medium.com/@techynilesh/xgboost-algorithm-explained-in-less-than-5-minutes-b561dcc1ccee

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Long short-term memory (LSTM) with Python – Data Science, Machine Learning, Deep Learning. (n.d.). Long Short-term Memory (LSTM) With Python – Data Science, Machine Learning, Deep Learning. https://www.alpha-quantum.com/blog/long-short-term-memory-lstm-with-python/long-short-term-memory-lstm-with-python/

How To Code RNN and LSTM Neural Networks in Python. (n.d.). How to Code RNN and LSTM Neural Networks in Python. https://www.nbshare.io/notebook/249468051/How-To-Code-RNN-and-LSTM-Neural-Networks-in-Python/

Singh, M. (2021, January 28). 4 Python libraries to detect English and Non-English language. Medium. https://towardsdatascience.com/4-python-libraries-to-detect-english-and-non-english-language-c82ad3efd430