Capstone Project

The News Classifier: Leveraging Machine Learning to Categorize Headlines

By: Theresia Novianne

Date: 16/02/2024

Table of Contents

Problem Statement

News articles generally update the readers on current events regarding a topic of interest, such as politics, climate, business, etc. On news websites, we see each article categorised by the topics. As of 2022, news articles are still being classified manually by the content managers (Singh, 2022). This is a problem as manually classifying news articles is time-consuming and decreases the productivity of employees. Ideally this process could be automated to increase the productivity of staff and in turn save costs for the business.

Classifying news headlines using machine learning has been addressed by other researchers. A project done by the Department of Information Science and Engineering on HuffPost and UCI dataset showed that machine learning models can classify news articles into its topics at an accuracy of up to 89.66% (Chhajerh, KVS, Meleet, & Murthy, 2021).

Business Context (Industry & Stakeholder)

For this project, the key internal stakeholders are news publication board members, investors, and employees. The key external stakeholders are news consumers and potential sponsors. The main objective of this project is to accurately classify news headlines into each category to align with the needs of our internal and external stakeholders. For our employees, automating news aggregation leads to time efficiency and increased productivity in other business operations. For our consumers, fine-tuning news headline classification enhances search functionality. For our potential sponsors, we can better define the target audience and refine our targeted advertising. For our board members and investors, we can expect a quick return on investment with strong long-term benefits. We predict to reduce our cost, improve business efficiency, and increase our profit margins from ad revenue (Turing, 2023).

This project is also relevant to other business contexts. Some other use cases include:

- Product categorization: E-commerce sites can use similar algorithms to classify their product by category (eg. Makeup products can be sorted into eyes, lips, and cheek products).
- Medical record classification: Hospital information system can sort each patient's medical results into the different category (eg. X-ray and MRI into the imaging category).

Business Question

Businesses generally look for ways to efficiently use time and reduce resource expenditure. Utilising machine learning models to categorise headlines enables employees to redistribute their productivity hours; improving business operations.The business question would be “Can we automate the news aggregation process?”.

Data

The data used for this project was obtained from HuffPost’s website and accessed through Kaggle, collected between 2012 to 2022 (Misra, 2022). The dataset contains 209,527 rows and 6 columns (link, headline, category, short_description, authors, date). The dataset had some content imbalances, where there are 200,000 headlines between 2012 and May 2018, and only 10,000 headlines between May 2018 and 2022. This dataset is available to download from Kaggle: https://www.kaggle.com/datasets/rmisra/news-category-dataset.

Data Question

The data question for this project would be “How accurately can machine learning and deep learning models classify news headlines into each category?” The main data points required to answer this question is the category, headline, and description features from the dataset.

Data Science Process

Data Analysis

The file for this dataset was loaded into Jupyter notebook as a dataframe. The dataframe had no nulls in the relevant columns (i.e. category, headline, and date columns).

The target variable in this project was the news category. There were 42 classes of news categories in the target with an imbalance in the number of headlines per category. The 42 categories were then sorted into a more manageable amount of 10 categories. The 10 news categories were as follows: Global, Entertainment, Family, Education, Sports, Politics, Environment, Lifestyle, Business, and Food.

For computing capability reasons, we had to reduce the size of the dataset. Otherwise, it would have taken a significant amount of time to conduct machine learning algorithms. We randomly sampled 2000 headlines for each category. The final working dataset now contains 20,000 headlines, instead of 210,000.

The first step in analysing is to preprocess the headline and shorten description. These two features were tokenized and had the stop words & special characters removed.

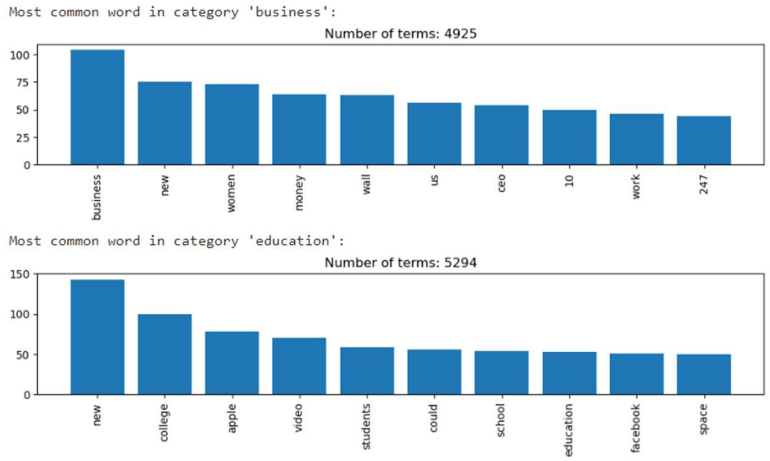


Figure 1. Bar chart of the 10 most common words in the category 'business' and 'education'.

Figure 2. Word cloud for the news category

Figure 3. Word cloud for the news category

Category distribution by the year the news article was released.

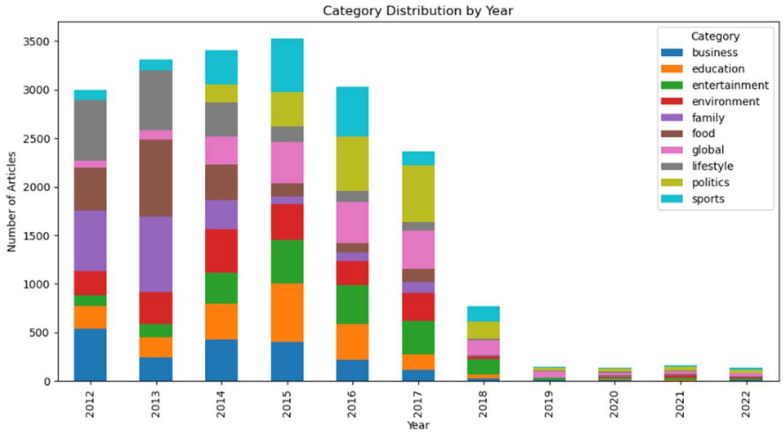


Figure 4. Category distribution by year. Note that because there are only 10,000 headlines in 2018-2022, the bar chart is also reflecting this.

Parts of Speech counts were also added to the dataset to help in analysing the trends in headlines.

	char_count	word_count	word_density	title_count	uppercase_count	adj_count	adv_count	noun_count	num_count	pron_count	propn_count	verb_count
2642	69	11	5.363636	2	0	0.0	0.0	2.0	2.0	1.0	4.0	1.0
195331	22	3	6.666667	9	0	0.0	0.0	1.0	0.0	1.0	0.0	1.0
62977	59	9	5.666667	14	1	0.0	0.0	1.0	0.0	0.0	6.0	0.0
18894	55	10	4.600000	6	0	0.0	0.0	2.0	0.0	0.0	2.0	1.0
80400	65	11	5.000000	6	2	1.0	0.0	1.0	0.0	0.0	6.0	1.0
73104	15	3	4.333333	4	0	0.0	0.0	0.0	0.0	0.0	3.0	0.0
181354	76	12	5.416667	8	1	0.0	0.0	1.0	0.0	0.0	7.0	1.0
137434	51	11	3.727273	1	0	1.0	0.0	1.0	1.0	0.0	5.0	0.0
48407	62	10	5.300000	13	2	0.0	0.0	1.0	0.0	0.0	5.0	1.0
147883	21	3	6.333333	8	0	1.0	0.0	0.0	0.0	0.0	0.0	1.0

Figure 5. A sample of 10 rows of the parts of speech that were

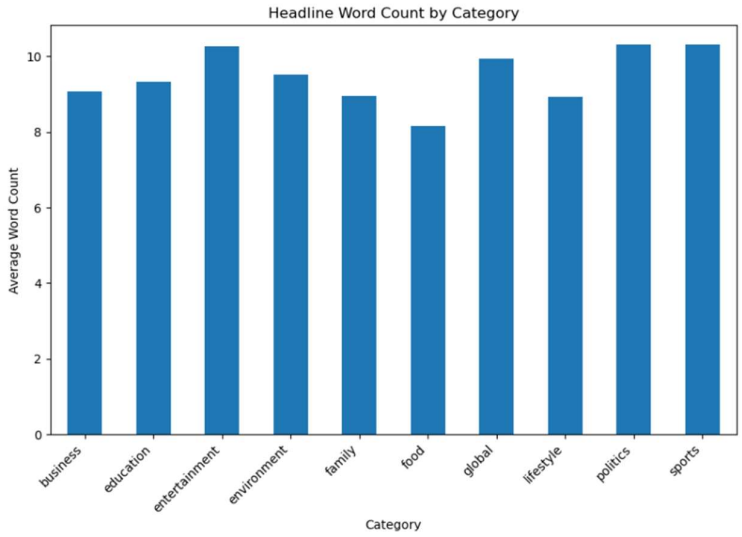


Figure 6. A bar chart of the average headline word count by category

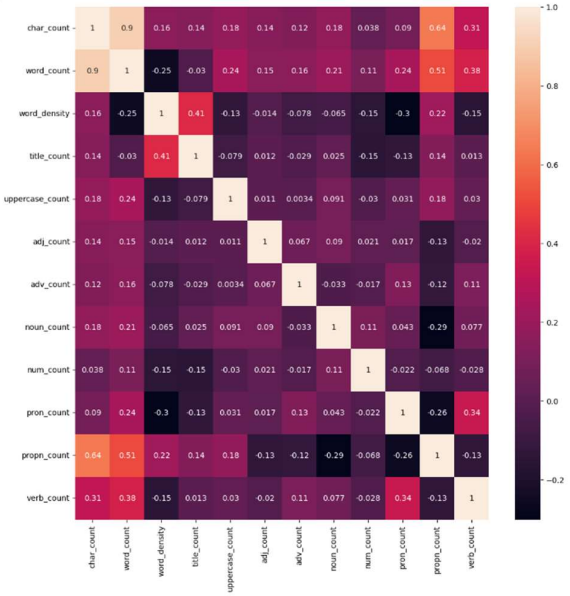


Figure 7. Parts of Speech Correlation Matrix.

For the final process in analysing the data, we exported the dataframe into a CSV file to use for the modelling Jupyter notebook and Google Colab notebook.

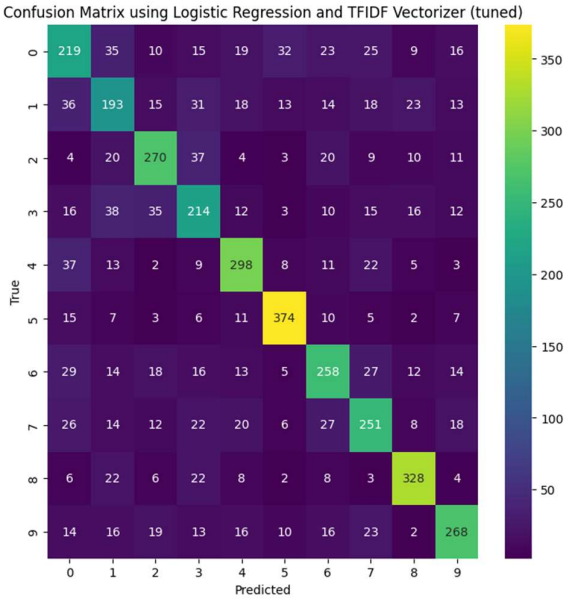
Modelling

The main features used for modelling were the tokenized headline & short description and the parts of speech. Parts of speech variables are as follows: word count, character count, word density, title word count, uppercase word count, adjective count, adverb count, noun count, number count, pronoun count, proper noun count, and verb count. The target variable is the category. The models selected were as follows: Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, and Gradient Boosting.

Each model was run 4 times: using Count Vectorizer, TF-IDF Vectorizer, and then tuned using RandomizedSearchCV with Count Vectorizer and TF-IDF Vectorizer. The Logistic Regression and Naïve Bayes model training and fitting took the least amount of time. The SVM and Gradient Boosting models took at least 30 minutes and up to 1 hour to train. The figures below show the confusion matrix and the classification report of the best version of the model.

Logistic Regression

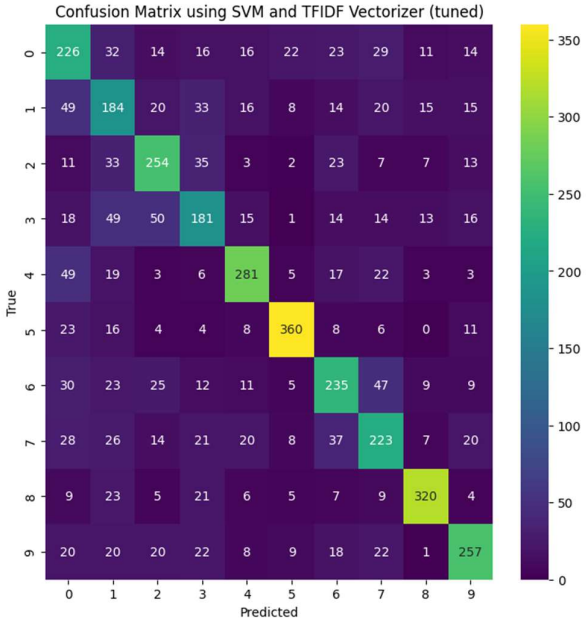
The best logistic regression model was tuned and paired with TF-IDF Vectorizer with an accuracy of 66.8%.



	precision	recall	f1-score	support
0	0.54	0.54	0.54	403
1	0.52	0.52	0.52	374
2	0.69	0.70	0.69	388
3	0.56	0.58	0.57	371
4	0.71	0.73	0.72	408
5	0.82	0.85	0.83	440
6	0.65	0.64	0.64	406
7	0.63	0.62	0.63	404
8	0.79	0.80	0.80	409
9	0.73	0.68	0.70	397
accuracy			0.67	4000
macro avg	0.66	0.66	0.66	4000
weighted avg	0.67	0.67	0.67	4000

Support Vector Machine (SVM)

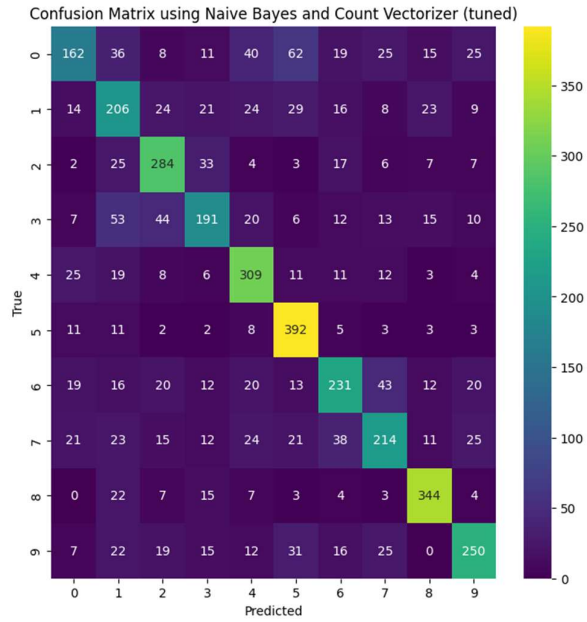
The best SVM model was tuned and paired with TF-IDF Vectorizer with an accuracy of 63%.



	precision	recall	f1-score	support
0	0.49	0.56	0.52	403
1	0.43	0.49	0.46	374
2	0.62	0.65	0.64	388
3	0.52	0.49	0.50	371
4	0.73	0.69	0.71	408
5	0.85	0.82	0.83	440
6	0.59	0.58	0.59	406
7	0.56	0.55	0.56	404
8	0.83	0.78	0.81	409
9	0.71	0.65	0.68	397
accuracy			0.63	4000
macro avg	0.63	0.63	0.63	4000
weighted avg	0.64	0.63	0.63	4000

Naïve Bayes

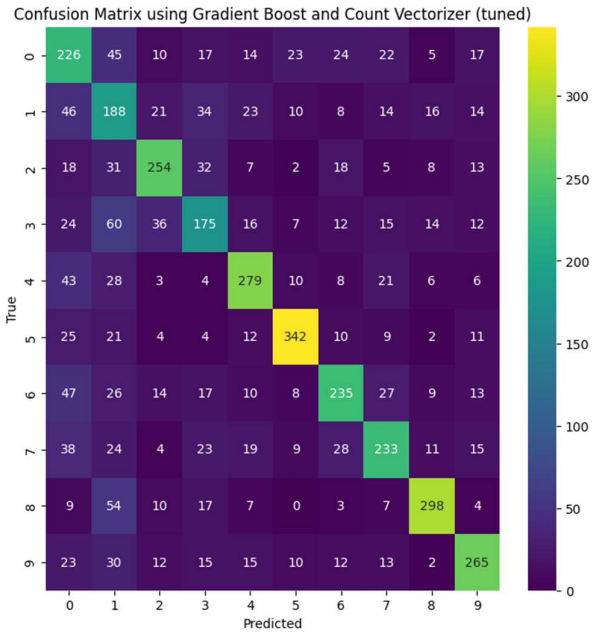
The best Naïve Bayes model was tuned and paired with Count Vectorizer with an accuracy of 64.6%.



	precision	recall	f1-score	support
0	0.60	0.40	0.48	403
1	0.48	0.55	0.51	374
2	0.66	0.73	0.69	388
3	0.60	0.51	0.55	371
4	0.66	0.76	0.71	408
5	0.69	0.89	0.78	440
6	0.63	0.57	0.60	406
7	0.61	0.53	0.57	404
8	0.79	0.84	0.82	409
9	0.70	0.63	0.66	397
accuracy			0.65	4000
macro avg	0.64	0.64	0.64	4000
weighted avg	0.64	0.65	0.64	4000

Gradient Boosting

The best gradient boosting model was tuned and paired with TF-IDF Vectorizer with an accuracy of 62.4%.



	precision	recall	f1-score	support
0	0.60	0.40	0.48	403
1	0.48	0.55	0.51	374
2	0.66	0.73	0.69	388
3	0.60	0.51	0.55	371
4	0.66	0.76	0.71	408
5	0.69	0.89	0.78	440
6	0.63	0.57	0.60	406
7	0.61	0.53	0.57	404
8	0.79	0.84	0.82	409
9	0.70	0.63	0.66	397
accuracy			0.65	4000
macro avg	0.64	0.64	0.64	4000
weighted avg	0.64	0.65	0.64	4000

Summary

Some models were improved significantly (from 28% to 60%) when hyperparameter tuning was conducted while others were only improved by a small percentage. Due to the computational power that RandomizedSearchCV requires, Google Colab Pro+ was also used to run the notebook.

	Count Vectorizer	TF-IDF Vectorizer
Logistic Regression	0.64275	0.66525
SVM	0.56800	0.28250
Naive Bayes	0.64375	0.64375
Boosting	0.46875	0.45300

Figure 8. Accuracy results before RandomizedSearchCV was performed.

	Count Vectorizer	TF-IDF Vectorizer
Logistic Regression	0.65175	0.66825
SVM	0.62375	0.63025
Naive Bayes	0.64575	0.64300
Boosting	0.62375	0.62425

Figure 9. Accuracy results after RandomizedSearchCV was performed.

Outcomes

The best model with 67% accuracy was the Logistic Regression with TF-IDF Vectorizer after it has been tuned. This confirms that the process of news categorization can be automated.

Implementation

The implementation of this model requires setting up necessary infrastructure to host and deploy the model. Another consideration is to ensure that the model can handle different volumes of data and traffic so that it can be used for new data (Calciano, 2023).

Data Answer

Using the HuffPost dataset, the tuned Logistic Regression with Count Vectorization model can categorise news headlines at 68% accuracy given 10 news categories.

Business Answer

The business question was answered. We can currently automate news article categorization with an accuracy of up to 68%. This is a significant improvement from the 10% baseline probability of categorizing a news article correctly into one of 10 categories.

End-to-End Solution

The end-to-end solution of this project requires some considerations of acquiring the necessary infrastructure to ensure that all steps from data acquisition to model deployment. This will provide users with an efficient and accurate for categorizing news contents (Turing, 2023).

Response to Stakeholders

The recommendation for this project is to investigate whether deep learning models could further improve the accuracy of the news categorisation. Although the current accuracy isn't up to par for business deployment, the initial results have been very promising.

Therefore, further investment to automate categorisation of news articles can decrease misclassification, refine targeted advertising, and improve reader engagement. This has a high potential to reduce cost, optimise business operations and increase consumer engagement.

Future Recommendations

As mentioned before, the current model needs to be further refined before it can be deployed for business operations. We suggest adding a deep learning model which would improve the accuracy of news categorisation beyond 67%.

News categorization can go beyond headlines. We could further categorise articles based on context and use labels/tags to better target the audience (eg. Articles related to Superbowl can be tagged with both sports and entertainment).

It would also be useful to have training data from other news sites apart from HuffPost so it can better classify worldwide headlines.

Bibliography

Calciano, D. (2023, May 17). *The Essential Hardware for Successful AI Implementation in Business: An In-depth Guide*. Retrieved from LinkedIn: https://www.linkedin.com/pulse/essential-hardware-successful-ai-implementation-guide-daniel-calciano/

Chhajerh, M. S., KVS, A., Meleet, P., & Murthy, D. (2021). Real TIme News Headlines Classification Using Machine Learning. *International Research Journal of Engineering and Technology (IRJET)*, 3296-3300.

Misra, R. (2022). *News Category Dataset*. Retrieved from Kaggle: https://rishabhmisra.github.io/publications/

Singh, D. (2022, August 22). *Text Classification of News Articles*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/

Turing, R. G. (2023, September 13). *AI Implementation Strategy Tips*. Retrieved from Turing: https://www.turing.com/blog/ai-implementation-strategy-tips/