
The News Classifier:
Leveraging Machine
Learning to Categorize
Headlines
Theresia Novianne
Capstone Project

About me

Theresia Novianne

- Bachelor of Medical Laboratory Science.
- 3 years working experience in the Laboratory as a Scientist.
- Aspiring Data Scientist.

Agenda

01 **Introduction**

- Business Context
- Project goals

02 **Data Exploration**

- Collect
- Clean
- Manipulation

03 **Analysis & Modelling**

- Visualisation
- Training and evaluation

04 **Conclusion**

- Conclusions
- Recommendations

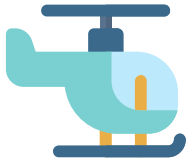
2-3 Million

News stories are released worldwide every 24 hours.

News Categorisation

- News articles are usually sorted into their topic categories (e.g. Politics, technology)
- Currently, some news organisations still have content managers that classify their content manually.
- This is a time-consuming task at the cost of the content manager's productivity.

Objectives



Business Goal

Automate news categorisation to increase employee productivity.

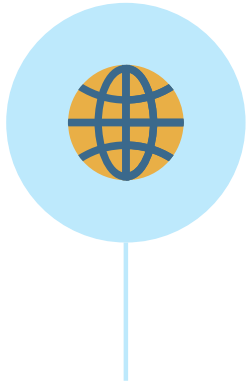
Project Goal

Predict the category of the news article by the headline and description.

Other Business Use Cases

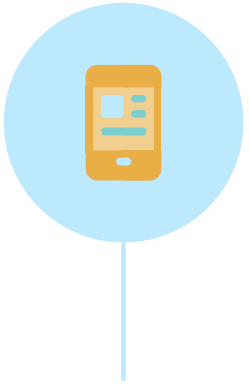
- Product categorization in E-commerce sites.
- Medical Record Classification.

Who will benefit?



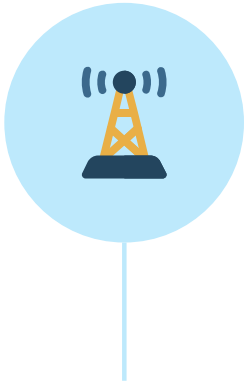
News Businesses

No manual categorisation required.



News Readers

Enhanced search funtionality to read their desired topic.



Sponsors

Easily advertise products related to topics.

Dataset

Source

- Kaggle
- Headlines extracted from HuffPost

Data

- 200,000 headlines between 2012-2018
- 10,000 headlines between 2018-2022

Columns

- Link
- Headline
- Category
- Short description
- Authors
- Date

Target

Category

Overview

Data Preprocessing

- Redefine the categories
- Subset the data
- Feature Engineering (Parts of Speech)

Models

- Logistic regression
- Naive Bayes
- Support Vector Machine
- Gradient Boosting

Evaluation

- Training and testing models
- Evaluate Accuracies

Tuning

- Perform Randomized Search Cross Validation to find the best parameters with each model
- Re-evaluate accuracies

Testing with New Data

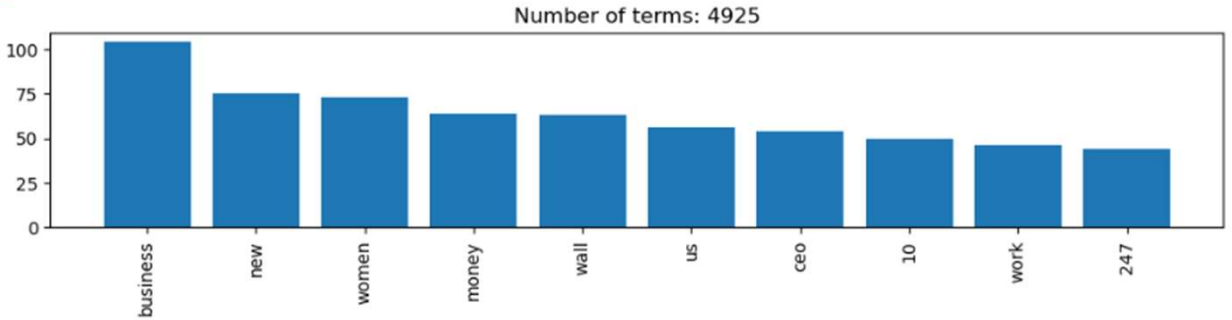
- Test the best model with some made-up headlines
- Conclusions and Recommend-ations

10 News Headline Categories

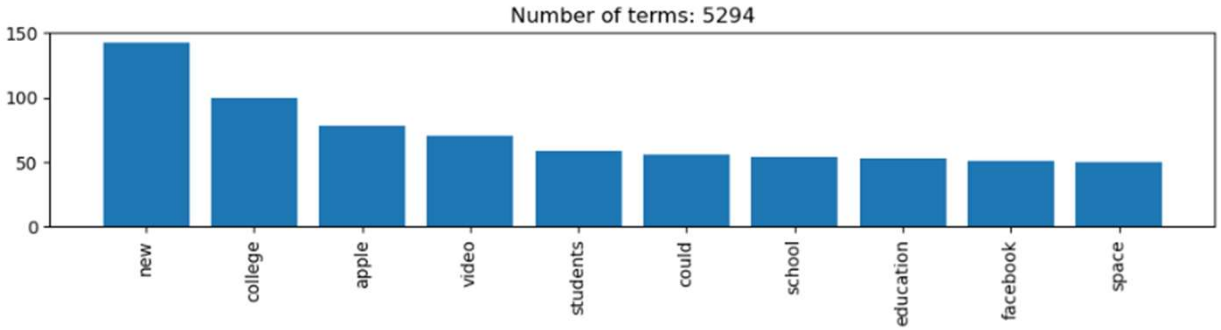
- Lifestyle
- Entertainment
- Politics
- Global
- Family
- Food
- Business
- Education
- Sports
- Environment

Most common words by category

Business

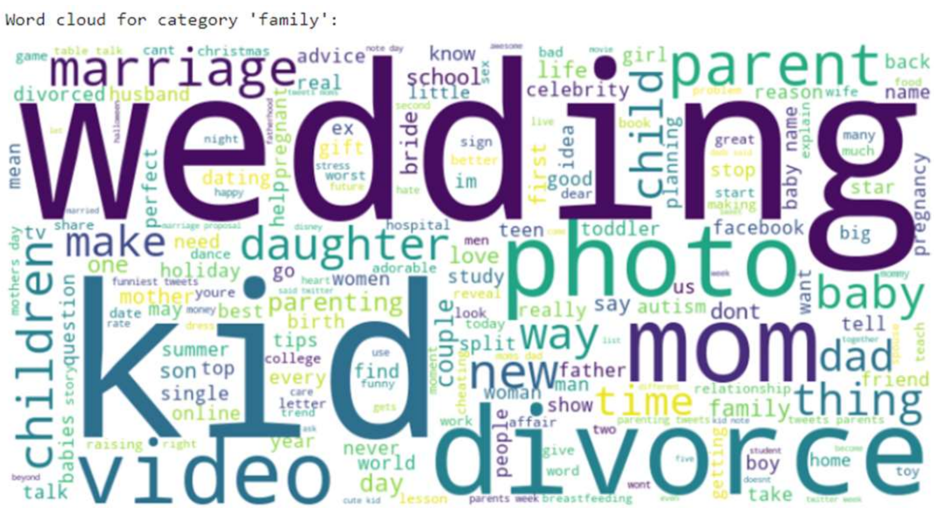


Education



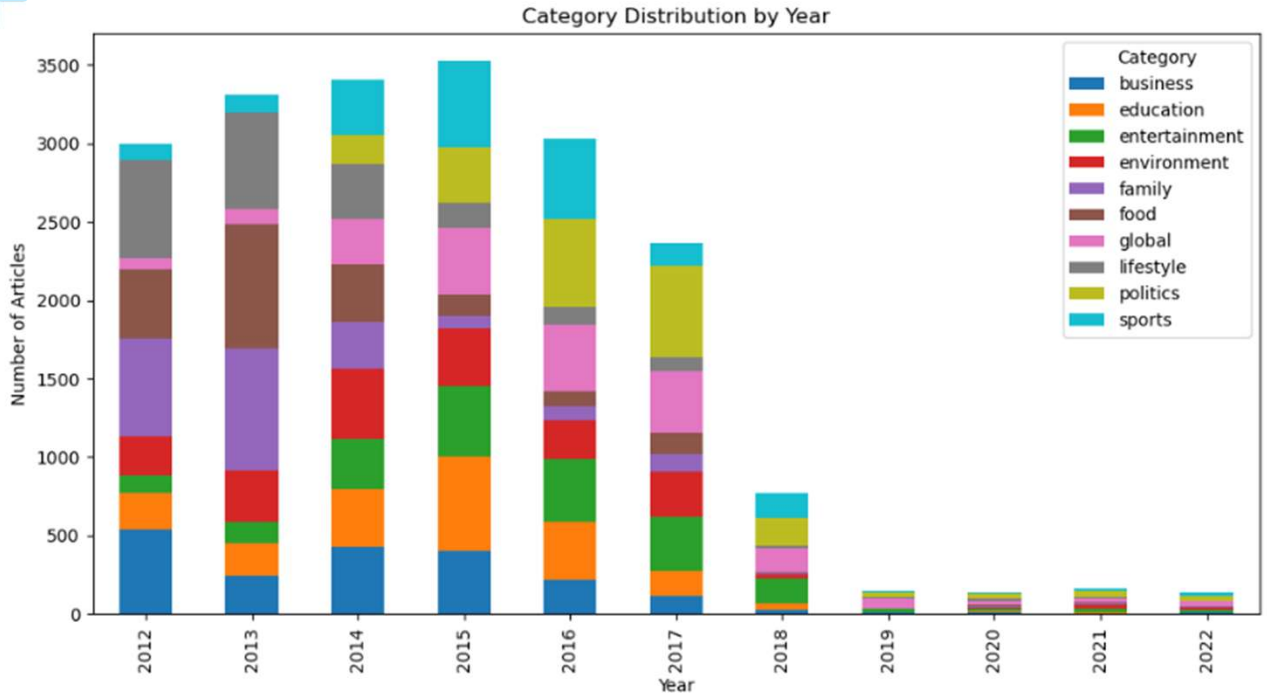
Most common words by category

Family

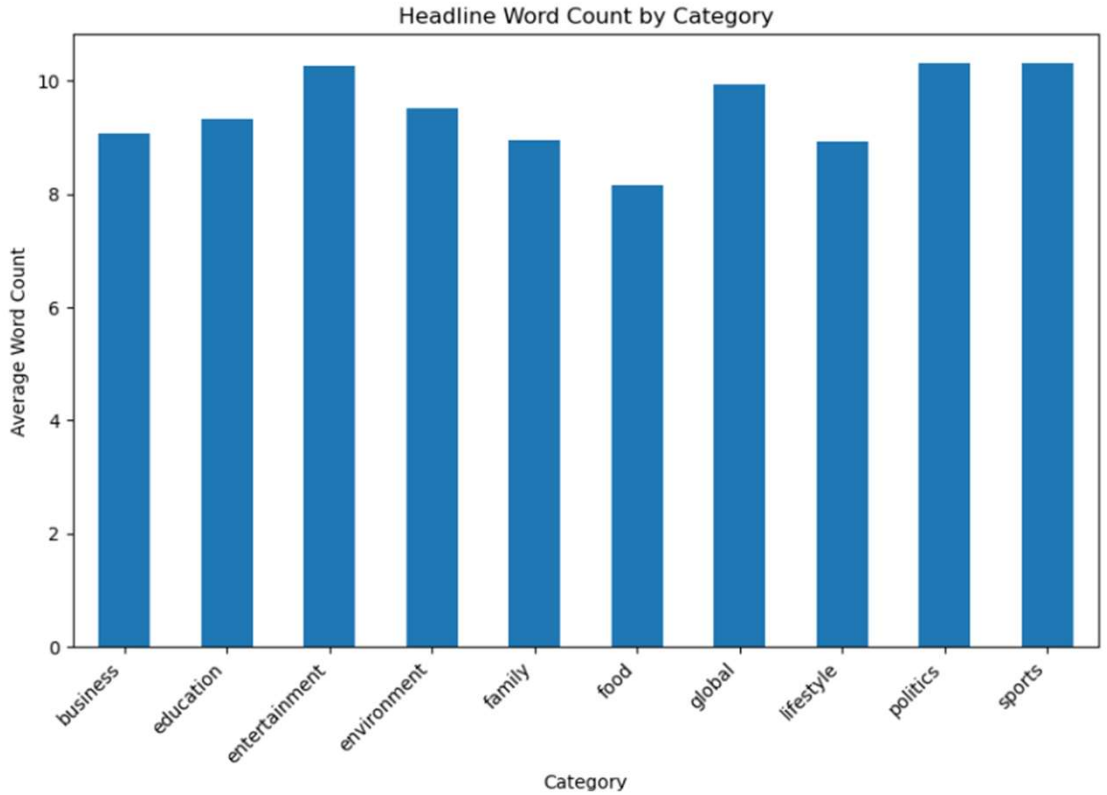


Food

Category Distribution by Year



Average headline word count by category

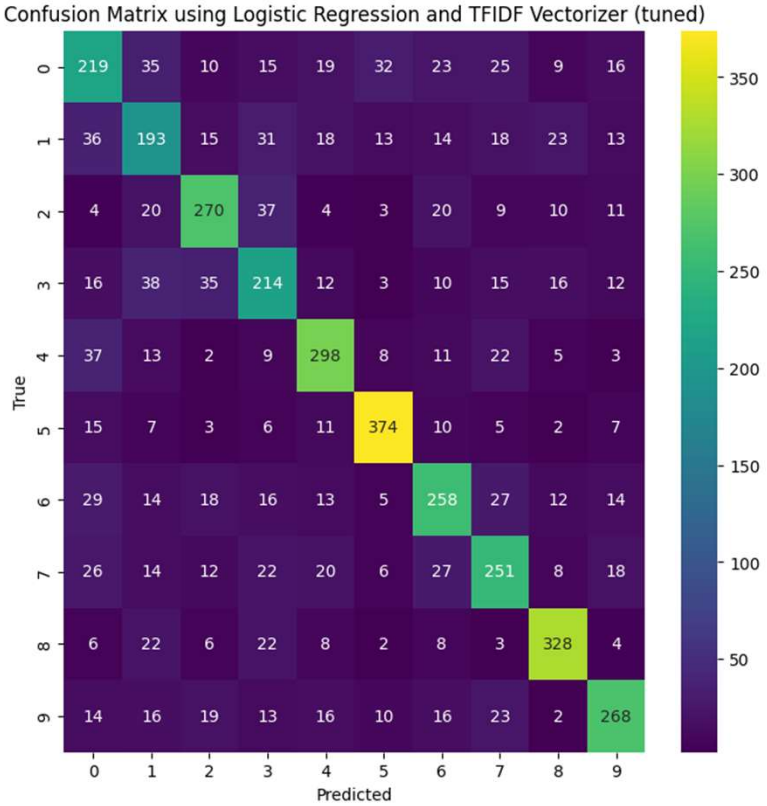


Modelling

Evaluation of Accuracies

	Logistic Regression	Support Vector Machines	Naïve Bayes	Gradient Boosting
Count Vectorizer (Baseline)	0.643	0.568	0.644	0.469
Count Vectorizer (Tuned)	0.652	0.624	0.646	0.624
TFIDF Vectorizer (Baseline)	0.665	0.283	0.644	0.453
TFIDF Vectorizer (Tuned)	**0.668**	0.630	0.643	0.624

Evaluation of the Best Model



0	Lifestyle
1	Entertainment
2	Politics
3	Global
4	Family
5	Food
6	Business
7	Education
8	Sports
9	Environment

	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

Testing the Best Model

Sample Predictions

Headline: A new study suggests drinking coffee may reduce the risk of heart disease
Description: Researchers have found evidence that drinking coffee may have cardiovascular benefits.
Predicted Category: lifestyle

Headline: The Feelers lead singer James Reid pleads guilty to refusing blood test
Description: The lead singer of one of the country's most prominent bands, The Feelers, has pleaded guilty to refusing a blood test.
Predicted Category: global

Headline: Department of Internal Affairs to bring case against SkyCity
Description: SkyCity is facing five separate civil proceedings brought by the Department of Internal Affairs. The legal action is related to the Anti-Money Laundering and Countering Financing of Terrorism Act. The casino operator could face a fine of up to $8 million
Predicted Category: politics

Consequences of Misclassification

Impact of credibility

The news business will be seen by readers as not credible leading to decreased readers.

Inappropriate content exposure

News articles that contains mature themes, if exposed to younger audiences, can lead to mistrust from parents.

Loss of Ad revenue

Sponsors may no longer want to associate with the publication as their ads are not reaching their target audience.

Conclusions

- Machine Learning models are up to 67% accurate at predicting the category.
- The best model at predicting headline categories was Logistic Regression with TF-IDF Vectorizer (tuned).

Recommendations

- Deep Learning model integration.
- News articles could have labels that can tag the article into multiple categories.
- Useful to have training data from other news sites apart from HuffPost.

Resources

- Slides: https://slidesgo.com/theme/global-news-agency#search-news&position-18&results-380&rs=search&rs=search
- Dataset: https://www.kaggle.com/datasets/rmisra/news-category-dataset
- News Article Statistics: https://earthweb.com/how-many-news-articles-are-published-every-day/
- News Classification: https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/
- News Headlines: https://www.stuff.co.nz/

Thank you!

Questions?
Comments?