

The background features a complex network of thin grey lines and dots, forming a web-like structure. Scattered throughout are various triangles of different sizes and orientations, some with solid dots at their vertices. The overall aesthetic is minimalist and technical, suggesting a focus on data or network analysis.

News Category Prediction

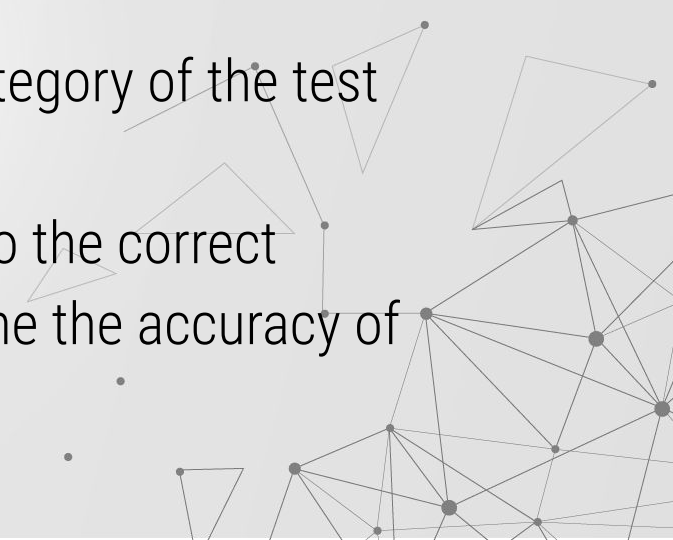
G14: Ting-Chih Chen, Huayu Liang, Zoe Zheng and Yilin Liu



01

Introduction

Introduction

- Open source data set from Microsoft MIND (Microsoft News Data set)
 - Our goal is to train four models to classify the news category from their titles and abstract.
 - Our different models can predict the category of the test news.
 - Using the predict results, we compare to the correct category within the data set and examine the accuracy of each model.
- 

02

Dataset



Dataset

- Our dataset is extracted from Microsoft News Dataset
- Totally, we have 101,527 news and 16 categories
- Each news has news title, abstract and category
- 80% for training
- 20% for testing

	NewID	Category	Title	Abstract	News	clean-News
0	N88753	lifestyle	The Brands Queen Elizabeth, Prince Charles, an...	Shop the notebooks, jackets, and more that the...	The Brands Queen Elizabeth, Prince Charles, an...	brand queen elizabeth prince charles prince ph...
1	N45436	news	Walmart Slashes Prices on Last-Generation iPads	Apple's new iPad releases bring big deals on l...	Walmart Slashes Prices on Last-Generation iPad...	walmart slash price last generation ipads appli...
2	N23144	health	50 Worst Habits For Belly Fat	These seemingly harmless habits are holding yo...	50 Worst Habits For Belly Fat These seemingly ...	worst habit belly fat seemingly harmless habi...
4	N93187	news	The Cost of Trump's Aid Freeze in the Trenches...	Lt. Ivan Molchanets peeked over a parapet of s...	The Cost of Trump's Aid Freeze in the Trenches...	cost trumpaid freeze trench ukrainewar lt ivan...

03

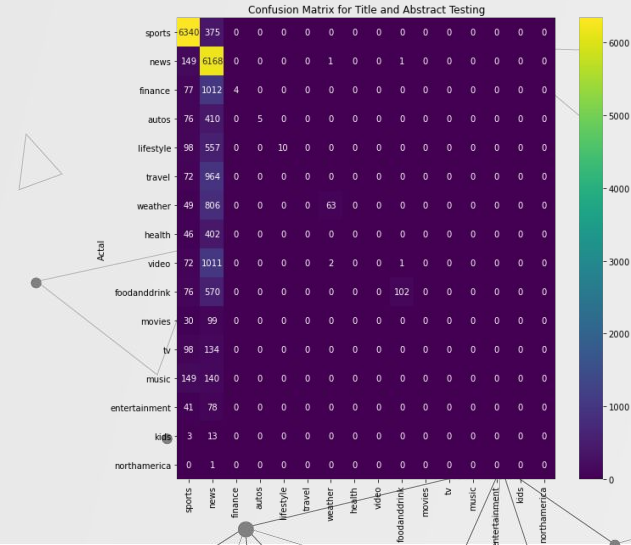
Methods



Naive Bayes

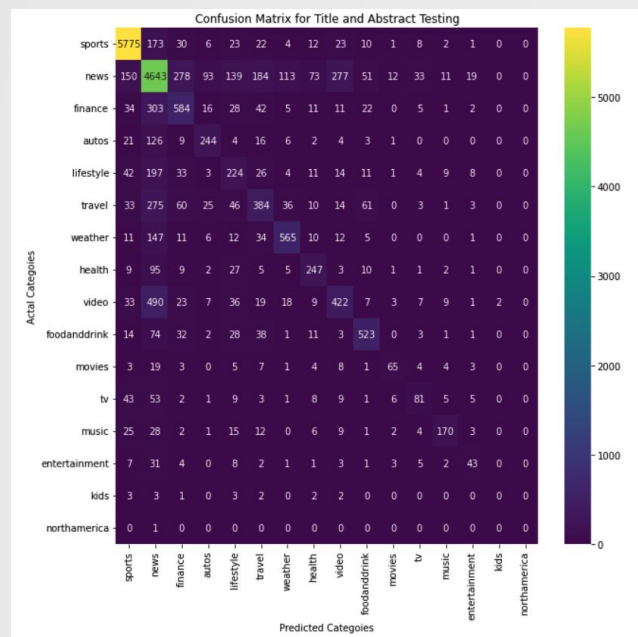
- NB is our baseline method
- First, we convert these news text to a matrix of token counts
- Second, we transform this count matrix to **tf** and **tf-idf** representations
- We use these representations matrix to train NB model
- NB accuracy:
 - **tf: 63%**
 - **tf-idf: 63%**

	precision	recall	f1-score	support
autos	1.00	0.01	0.02	491
entertainment	0.00	0.00	0.00	119
finance	1.00	0.00	0.01	1093
foodanddrink	0.98	0.14	0.24	748
health	0.00	0.00	0.00	448
kids	0.00	0.00	0.00	16
lifestyle	1.00	0.02	0.03	665
movies	0.00	0.00	0.00	129
music	0.00	0.00	0.00	289
news	0.48	0.98	0.65	6319
northamerica	0.00	0.00	0.00	1
sports	0.86	0.94	0.90	6715
travel	0.00	0.00	0.00	1036
tv	0.00	0.00	0.00	232
video	0.00	0.00	0.00	1086
weather	0.95	0.07	0.13	918
accuracy			0.63	20305
macro avg	0.39	0.13	0.12	20305
weighted avg	0.62	0.63	0.52	20305



Logistic Regression

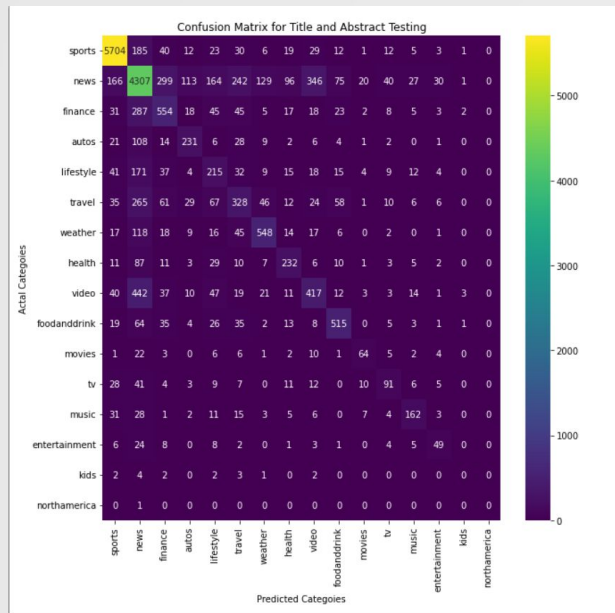
- The second algorithm is Logistic Regression
- Convert the collection of news text to a matrix of token counts (CountVectorizer)
- Then convert previous matrix to a matrix of TF and TF-IDF features (TfidfTransformer)
- Using the TF and TF-IDF features as the input to train the Logistic Regression model
- Logistic Regression accuracy on testing:
TF **73%**
TF-IDF **75%**



	precision	recall	f1-score	support
autos	0.71	0.50	0.58	436
entertainment	0.69	0.32	0.44	111
finance	0.64	0.50	0.56	1064
foodanddrink	0.74	0.68	0.71	731
health	0.71	0.54	0.61	417
kids	0.00	0.00	0.00	16
lifestyle	0.43	0.33	0.37	587
movies	0.71	0.33	0.45	127
music	0.83	0.50	0.62	278
news	0.66	0.87	0.75	6076
northamerica	0.00	0.00	0.00	1
sports	0.92	0.96	0.94	6090
travel	0.61	0.35	0.45	951
tv	0.66	0.18	0.28	227
video	0.68	0.29	0.40	1086
weather	0.79	0.68	0.73	814
accuracy			0.75	19012
macro avg	0.61	0.44	0.49	19012
weighted avg	0.75	0.75	0.73	19012

SVM

- The third algorithm is SVM
- Convert the collection of news text to a matrix of token counts (CountVectorizer)
- Then convert previous matrix to a matrix of TF-IDF features (TfidfTransformer)
- Using the TF / TF-IDF features as the input to train the SVM model
- SVM accuracy on testing:
tf 71% / tf-idf 73%



	precision	recall	f1-score	support
autos	0.53	0.53	0.53	436
entertainment	0.43	0.44	0.44	111
finance	0.49	0.52	0.51	1064
foodanddrink	0.70	0.70	0.70	731
games	0.00	0.00	0.00	0
health	0.52	0.56	0.54	417
kids	0.00	0.00	0.00	16
lifestyle	0.32	0.37	0.34	587
middleeast	0.00	0.00	0.00	0
movies	0.56	0.50	0.53	127
music	0.64	0.58	0.61	278
news	0.70	0.71	0.70	6076
northamerica	0.00	0.00	0.00	1
sports	0.93	0.94	0.93	6090
travel	0.39	0.34	0.36	951
tv	0.46	0.40	0.43	227
video	0.45	0.38	0.42	1086
weather	0.70	0.67	0.68	814
accuracy			0.71	19012
macro avg	0.43	0.43	0.43	19012
weighted avg	0.70	0.71	0.70	19012

Neural Network

Dense layer	1 layer	2 layers
50	0.72	0.72
128	0.73	0.71
256	0.73	N/A

We use the Keras Python library to perform multi-class classification task on news category prediction using Neural Network. The input sequence is a BoW matrix with each row represent each news and each column represent each words in that news. We built a fully connected Neural Network with one dense layer and a Softmax layer at the end to generated the category probabilities.

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	14556544
dense_1 (Dense)	(None, 18)	2322

Total params: 14,558,866

Trainable params: 14,558,866

Non-trainable params: 0

relu
softmax

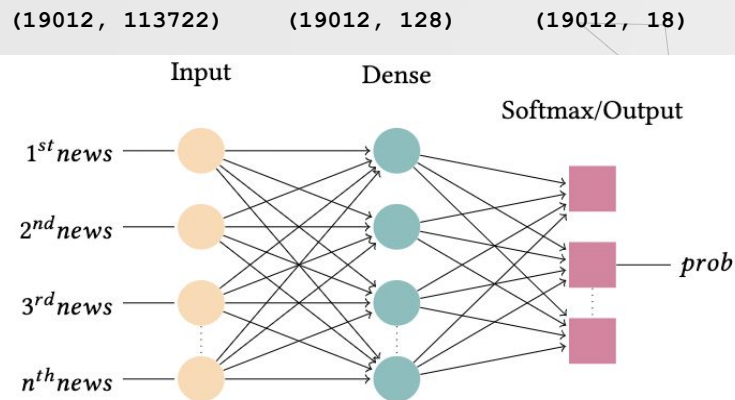
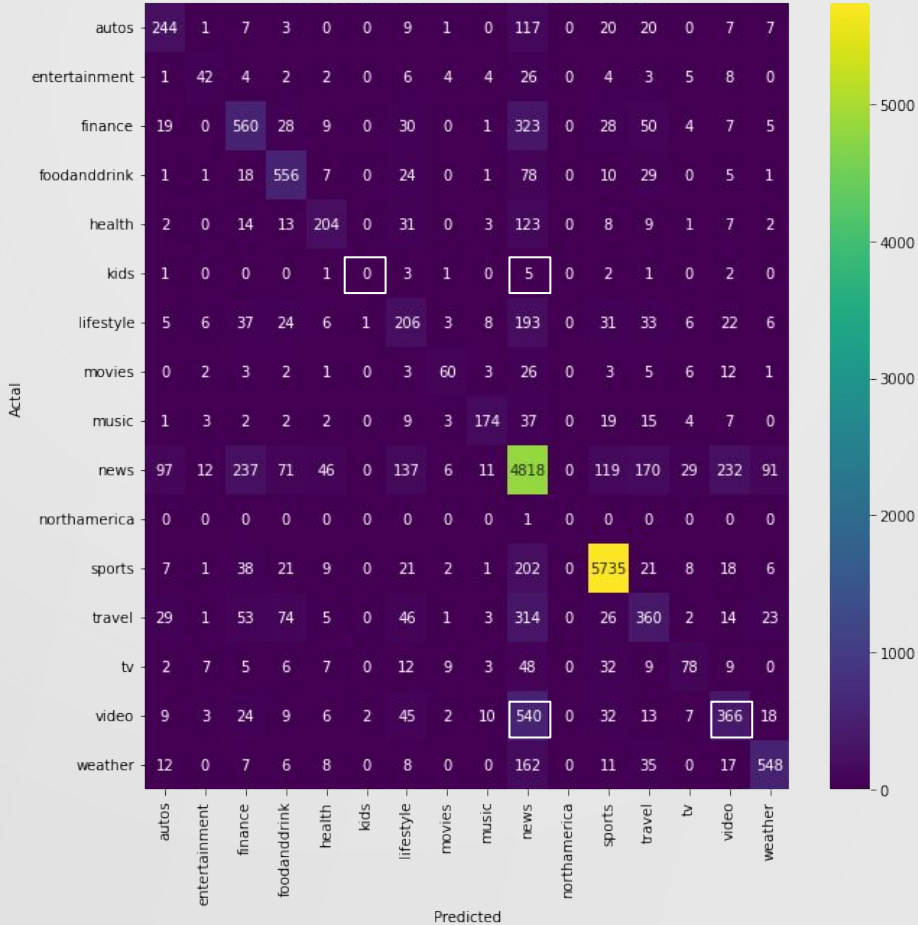


Figure 4. The architecture of Neural Network Model

Neural Network

Confusion Matrix for Title and Abstract Testing



Test Accuracy: 73.379970

From the confusion matrix, we can see our model has predictive ability by successfully classifying the majority of news in most category.

04

Results



Result

	TF Accuracy	TF-IDF Accuracy
NB	0.63	0.63
SVM	0.71	0.73
LR	0.73	0.75
NN	0.73	N/A

05

Contributions



Contribution

Ting-Chih Chen	Data pre-processing	Naive Bayes	Evaluation
Huayu Liang		Logistic Regression	Evaluation
Zoe Zheng	Data pre-processing	Neural Network	Evaluation
Yilin Liu		SVM	Evaluation

