NEWS TITLE CLASSIFICATION REPORT

Phu Quoc Trung
0363587268
phuquoctrung2003@gmail.com

DATASET

Dataset Summary:

- Dataset contains references to news web pages gathered from an online aggregator between March 10 and August 10, 2014.
- It encompasses 422,937 news pages categorized into business, science and technology, entertainment, and health.
- The dataset includes clusters of similar news stories and pairs of URLs representing browsing sessions.
- This report will focus solely on the file "newsCorpora.csv" which contains news pages, structured with columns including ID, TITLE, URL, PUBLISHER, CATEGORY, STORY, HOSTNAME, and TIMESTAMP.
- For the purpose of this report, only the TITLE, CATEGORY, and PUBLISHER columns will be utilized.

ID	TITLE	URL	PUBLISHER	CATEGORY	STORY	HOSTNAME	TIMESTAMP
1	Fed official says weak data caused by weather,...	http://www.latimes.com/business/money/la-fi-mo...	Los Angeles Times	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.latimes.com	1394470370698
2	Fed's Charles Plosser sees high bar for change...	http://www.livemint.com/Politics/H2EvwJSK2VE6O...	Livemint	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.livemint.com	1394470371207
3	US open: Stocks fall after Fed official hints ...	http://www.ifamagazine.com/news/us-open-stocks...	IFA Magazine	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.ifamagazine.com	1394470371550
4	Fed risks falling 'behind the curve', Charles ...	http://www.ifamagazine.com/news/fed-risks-fall...	IFA Magazine	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.ifamagazine.com	1394470371793
...	...	...	...	...	...	...	...

DATASET

Dataset Construction:

- First, retrieve records with PUBLISHER values of "Reuters", "Huffington Post", "Businessweek", "Contactmusic.com", and "Daily Mail“, we will focus on these publishers articles
- Then, select only the CATEGORY and TITLE columns.
- Clean TITLE columns by removing special symbols, whitespace errors, and lower all characters.
- Shuffle the dataset and then divide the extracted examples into three sets: 80% for training, 10% for validation, and 10% for evaluation.

CATEGORY	TITLE
b	ohio senate panel approves compromise on tesla…
b	asian stocks drop after worst weekly loss sinc…
e	beyonce changes lyrics to resentment internet …
m	artificial pancreas could help stem the diabetes e…
t	toyota admits it misled the public about multiple…
…	…

SET	SIZE
train set	10684
valid set	1336
Test set	1336

DATA EXPLORATION

	Number of samples for each category				Number of words in samples		
SET	**b**	**e**	**t**	**m**	**Average**	**Maximum**	**Minimum**
train	4500	4223	1229	732	10.77	19	2
valid	556	543	153	84	10.84	18	2
Test	571	528	143	94	10.76	18	3

Model Selection:

- Given the average word count of 10 and a maximum of 19, a simple model like LSTM seems appropriate.
- However, due to limited data and significant class imbalance, considering fine-tune a pre-trained models like BERT for enhanced performance is advisable.

Approach:

- I'll use LSTM due to lack of hardware resource for deployment, but will also include a notebook on the pre-trained BERT model I've fine-tuned.

DATA EXPLORATION (ADDITIONAL)

Topic Modeling Approach:

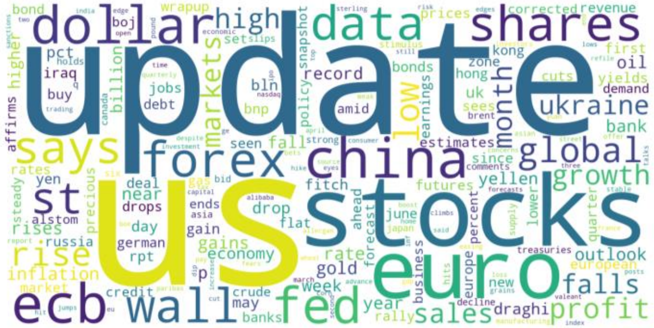
- In this section, I'll employ Labeled LDA to generate topic-related vocabulary sets along with their associated weights.
- This aims to evaluate which words significantly influence the classification of sentences into specific topics.
- I trained the model with data from all 3 sets, trained for 150 iteration (until almost converged).

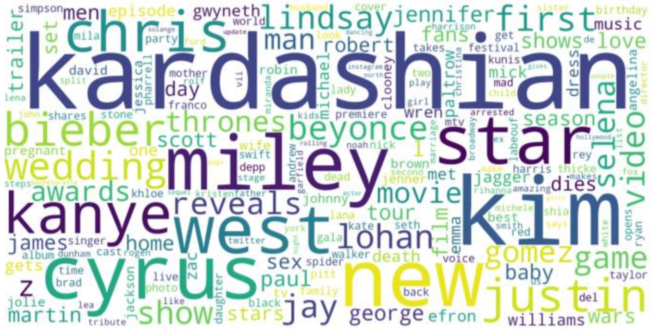
Credits:

- I'll utilize the code from https://github.com/JoeZJH/Labeled-LDA-Python, which is an implementation of the paper titled "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora" by the Computer Science Department at Stanford University.

DATA EXPLORATION (ADDITIONAL)

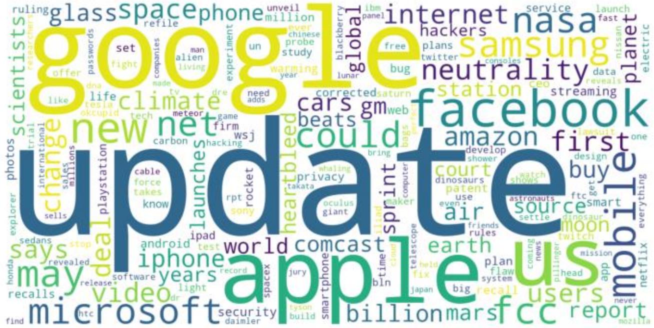
Bussiness

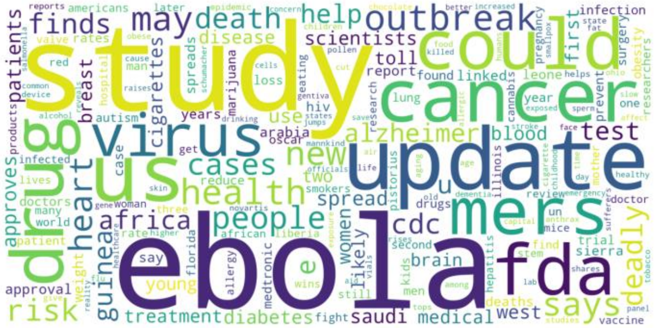




Entertainment

Science and Technology





Health

- The words influencing the titles of the articles are quite understandable and are significantly influenced by the social context at that time.
- This part could be more beneficial for fine-tuning and developing more sophisticated models for various purposes. However, to keep the assignment simple, I won't apply it here.

LSTM FOR MULTICLASS CLASSIFICATION

Data Preparation:

- **Reading Data:** Read in the training, validation, and test data from files.
- **Label Encoding:** Encode the categorical labels into numerical labels.
- **Balancing Dataset:** Create a balanced validation dataset by randomly sampling an equal number of samples for each class in the validation set. This step ensures an unbiased evaluation of the model's performance across all classes.
- **Tokenizer Creation:** Fit a tokenizer on the combined text data from all datasets and tokenize the text data.
- **Data Transform:** Convert the text data into sequences, pad the sequences to a fixed length, and convert labels to one-hot encoded format, in order to feed this to the model.

LSTM FOR MULTICLASS CLASSIFICATION

Model Structure:

- **An Input Layer:** Input layer.
- **An Embedding Layer:** input dimension = vocabulary_size and output dimension = 200.
- **A Spatial Dropout Layer:** Spatial Dropout layer with dropout rate = dropout (to be tuned during Bayesian optimization).
- **LSTM Layers:**
  - Multiple LSTM layers with num_units units, each with dropout rate = dropout (to be tuned during Bayesian optimization).
  - The last LSTM layer does not return sequences.
  - Number of layers will be tuned during Bayesian optimization.
- **Output Layer:** Dense layer with 4 units (for 4 classes) and softmax activation function.
- **Compilation:** Using Adam optimizer and categorical cross-entropy loss function.

LSTM FOR MULTICLASS CLASSIFICATION

Model Training:

- **epochs and batch_size:** Trained with a batch size of 256 and early stopping based on validation loss.
- **class_weight:** Dict of weight applied to each class to address class imbalance during model training. Calculated base on number of samples in each class.
- **validation_data:** Here I use the balanced valid set to validate the model.

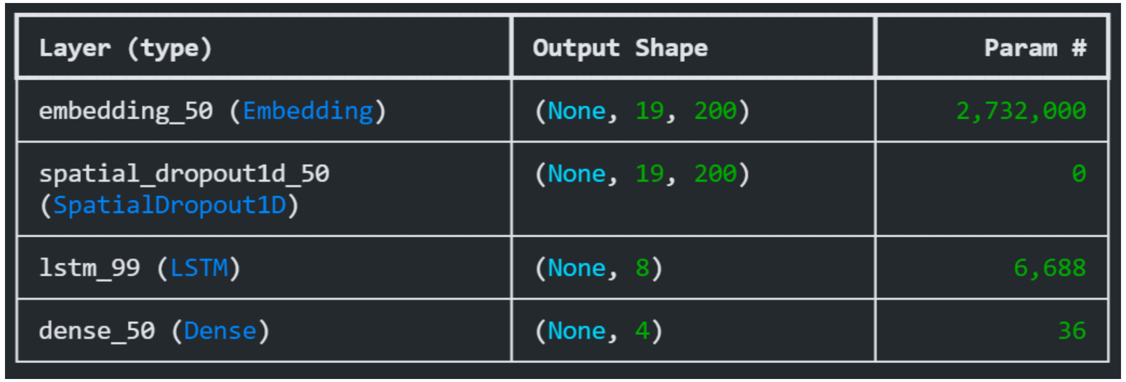
Hyper Parameter Tuning:

- Here I use **Bayesian optimization** to tune num_layers, num_units and dropout, with max evaluation count = 50.
- **num_layers:** Number of LSTM layers, an integer in between 1 and 4.
- **num_units:** Number of unit in each LSTM layer, an integer in between 8 and 32, with a step size of 8.
- **dropout:** The dropout rate for spatial dropout and LSTM layers, a float between 0.1 and 0.8.

LSTM FOR MULTICLASS CLASSIFICATION

Result:

- After tuning and training, we have the model with 1 LSTM layer, which have 8 units, and dropout rate = 0.55.



```
┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━┓
┃ Layer (type)                   ┃ Output Shape           ┃      Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━┩
│ embedding_50 (Embedding)       │ (None, 19, 200)        │    2,732,000 │
├────────────────────────────────┼────────────────────────┼──────────────┤
│ spatial_dropout1d_50           │ (None, 19, 200)        │            0 │
│ (SpatialDropout1D)             │                        │              │
├────────────────────────────────┼────────────────────────┼──────────────┤
│ lstm_99 (LSTM)                 │ (None, 8)              │        6,688 │
├────────────────────────────────┼────────────────────────┼──────────────┤
│ dense_50 (Dense)               │ (None, 4)              │           36 │
└────────────────────────────────┴────────────────────────┴──────────────┘
```

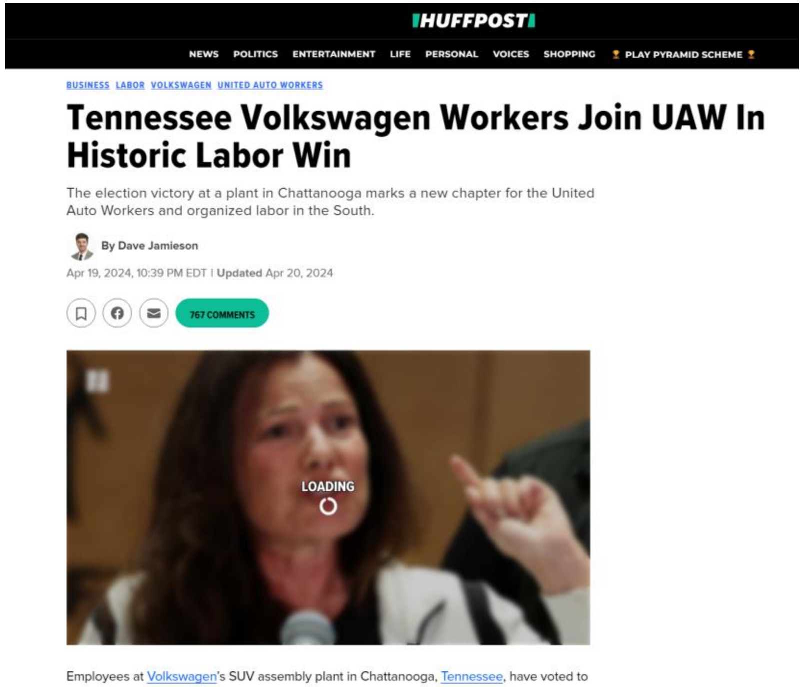
- Evaluate on test set show strong performance on the "Entertainment" and "Business" labels, with high Precision, Recall, and F1-score values. However, performance on the "Science and Technology" and "Health" labels is slightly lower, especially in terms of Recall and F1-score. The macro-average indicates decent overall performance across all labels, with an F1-score of 0.865. Nonetheless, improving performance on the "Science and Technology" and "Health" labels could enhance the model's generalization capability.

	Precision	Recall	F1-score
Entertainment	0.973	0.943	0.958
Business	0.954	0.946	0.950
Science and Technology	0.761	0.825	0.792
Health	0.728	0.798	0.761
Macro-average	0.854	0.878	0.865

MODEL DEPLOYMENT

- Model implementation into an API using FastAPI, structured as a simple MVC framework.
- The model, named NewsTitleClassificationModel, loads the checkpoint of the trained model and performs prediction tasks, as well as listing labels.
- A service named NewsTitleClassificationService, which serves as a service layer responsible for handling operations related to news title classification.
- A router named newsTitleClassificationRouter, which defines the API routes and their corresponding handlers for interacting with the NewsTitleClassificationService.
- API run with Base URL: http://localhost:2005 with end-point:
    - */list_label* (GET): get the list of labels available with the model.
    - */classify* (POST): get the label and it associated confident probability of the input text.
- A simple web page front-end is included to demo the api.

MODEL DEPLOYMENT



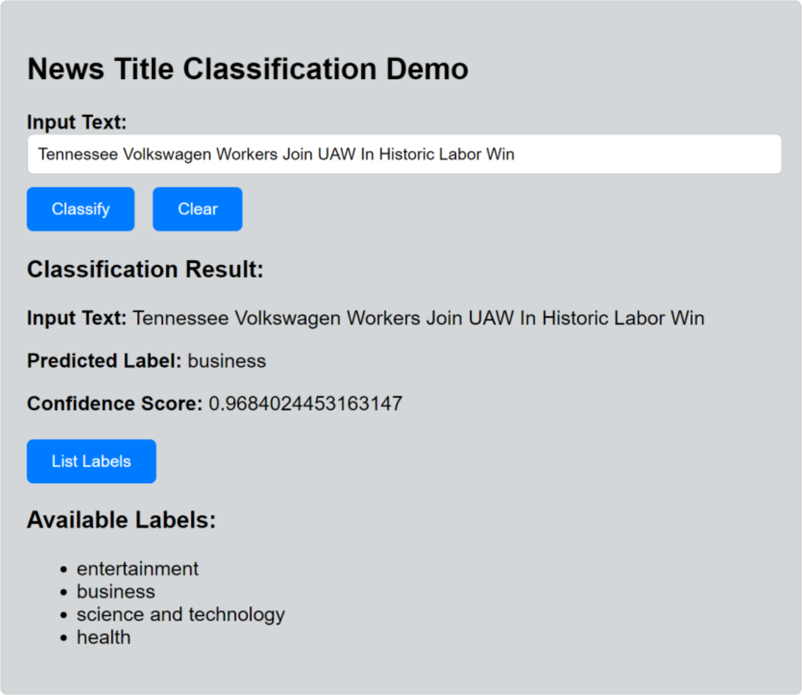
Tennessee Volkswagen Workers Join UAW In Historic Labor Win

The election victory at a plant in Chattanooga marks a new chapter for the United Auto Workers and organized labor in the South.

By Dave Jamieson

Apr 19, 2024, 10:39 PM EDT | **Updated** Apr 20, 2024

Employees at Volkswagen's SUV assembly plant in Chattanooga, Tennessee, have voted to



News Title Classification Demo

Input Text:

Tennessee Volkswagen Workers Join UAW In Historic Labor Win

Classification Result:

Input Text: Tennessee Volkswagen Workers Join UAW In Historic Labor Win

Predicted Label: business

Confidence Score: 0.9684024453163147

Available Labels:

- entertainment
- business
- science and technology
- health

Classification on the title of a recent article from the Huffington Post shows quite good results with a correct classification and a confidence score of 0.97.

THANK YOU AND LOOKING FORWARD TO HEARING FROM YOU!

Phu Quoc Trung
0363587268
phuquoctrung2003@gmail.com