DEEP NEWS – USING ADVANCED NLP TECHNIQUES TO CATEGORIZE NEWS ARTICLES

AIT PROJECT
FINAL PRESENTATION

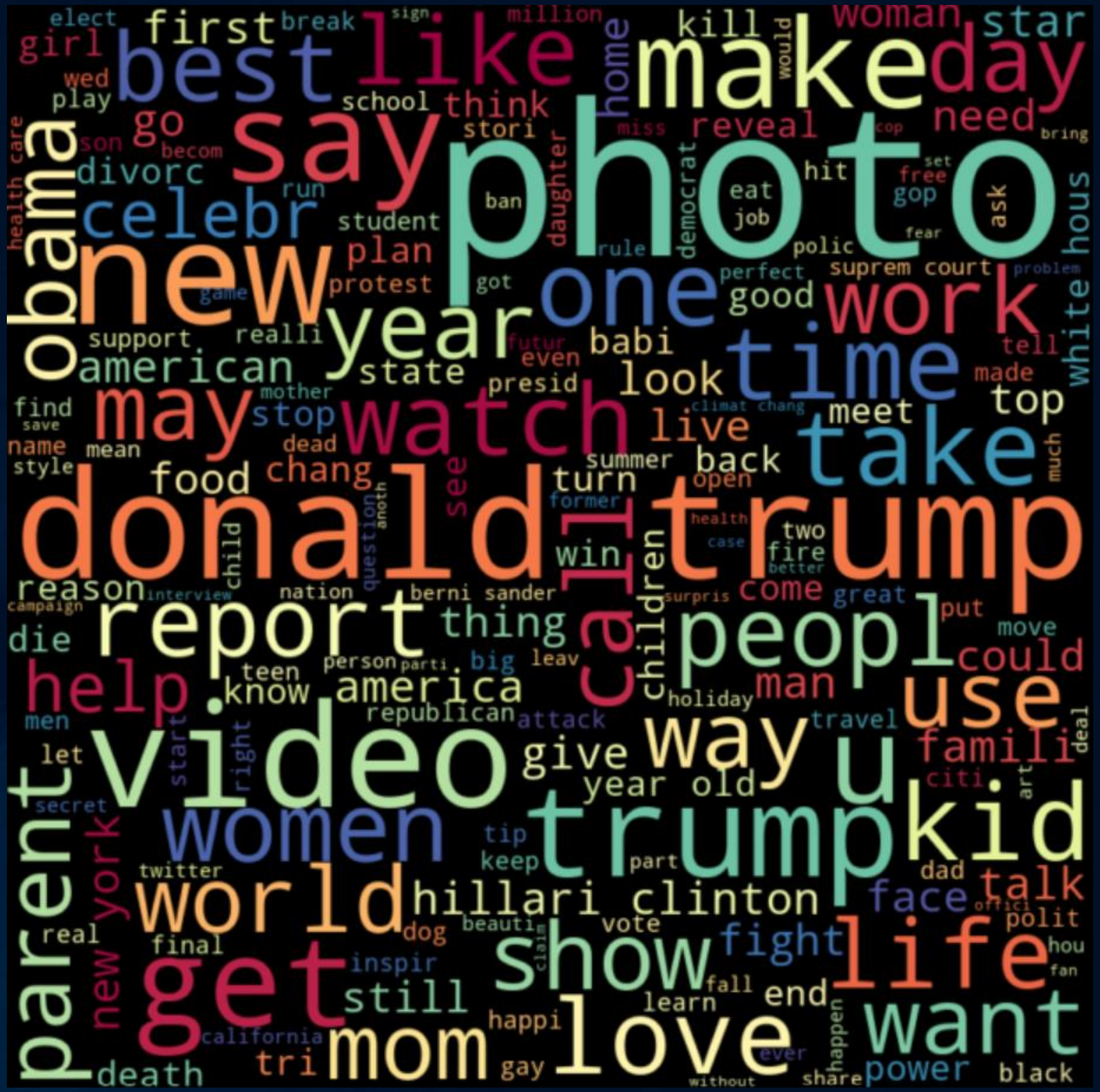
Presented by

Alan Gaugler

U885853

University of Canberra

October 26, 2023



INTRODUCTION

Why NLP?

Natural Language Processing (NLP) is a very important branch of AI with many practical applications that many of us use every day.

Summary of Literature Review

- Many articles were reviewed on the topic of NLP.
- The News Category Dataset was chosen due to its size and challenges.
- A good understanding of what is involved in an NLP classification problem.
- A good understanding of what vectorization and word embedding techniques are required.
- A good understanding of the various traditional and advanced deep learning models that are options for an NLP project.

DATASET – NEWS CATEGORY DATASET

- From Kaggle.
- HuffPost – a progressive and popular American news website.
- 209,527 news headlines from 2012 to 2022.
- 6 Features.
- Target Variable is ‘Category’.
- 42 Classes – Categories of news articles.
- ‘Headline’ and ‘Short Description’ merged
- Average of 29 words per article.

	link	headline	category	short_description	authors	date
0	https://www.huffpost.com/entry/covid-boosters-...	Over 4 Million Americans Roll Up Sleeves For O...	U.S. NEWS	Health experts said it is too early to predict...	Carla K. Johnson, AP	2022-09-23
1	https://www.huffpost.com/entry/american-airlin...	American Airlines Flyer Charged, Banned For Li...	U.S. NEWS	He was subdued by passengers and crew when he ...	Mary Papenfuss	2022-09-23
2	https://www.huffpost.com/entry/funniest-tweets...	23 Of The Funniest Tweets About Cats And Dogs ...	COMEDY	"Until you have a dog you don't understand wha...	Elyse Wanshel	2022-09-23
3	https://www.huffpost.com/entry/funniest-parent...	The Funniest Tweets From Parents This Week (Se...	PARENTING	"Accidentally put grown-up toothpaste on my to...	Caroline Bologna	2022-09-23
4	https://www.huffpost.com/entry/amy-cooper-lose...	Woman Who Called Cops On Black Bird-Watcher Lo...	U.S. NEWS	Amy Cooper accused investment firm Franklin Te...	Nina Golgowski	2022-09-22

Figure 1 – The Header of the News Category Dataset

EXPLORATORY DATA ANALYSIS

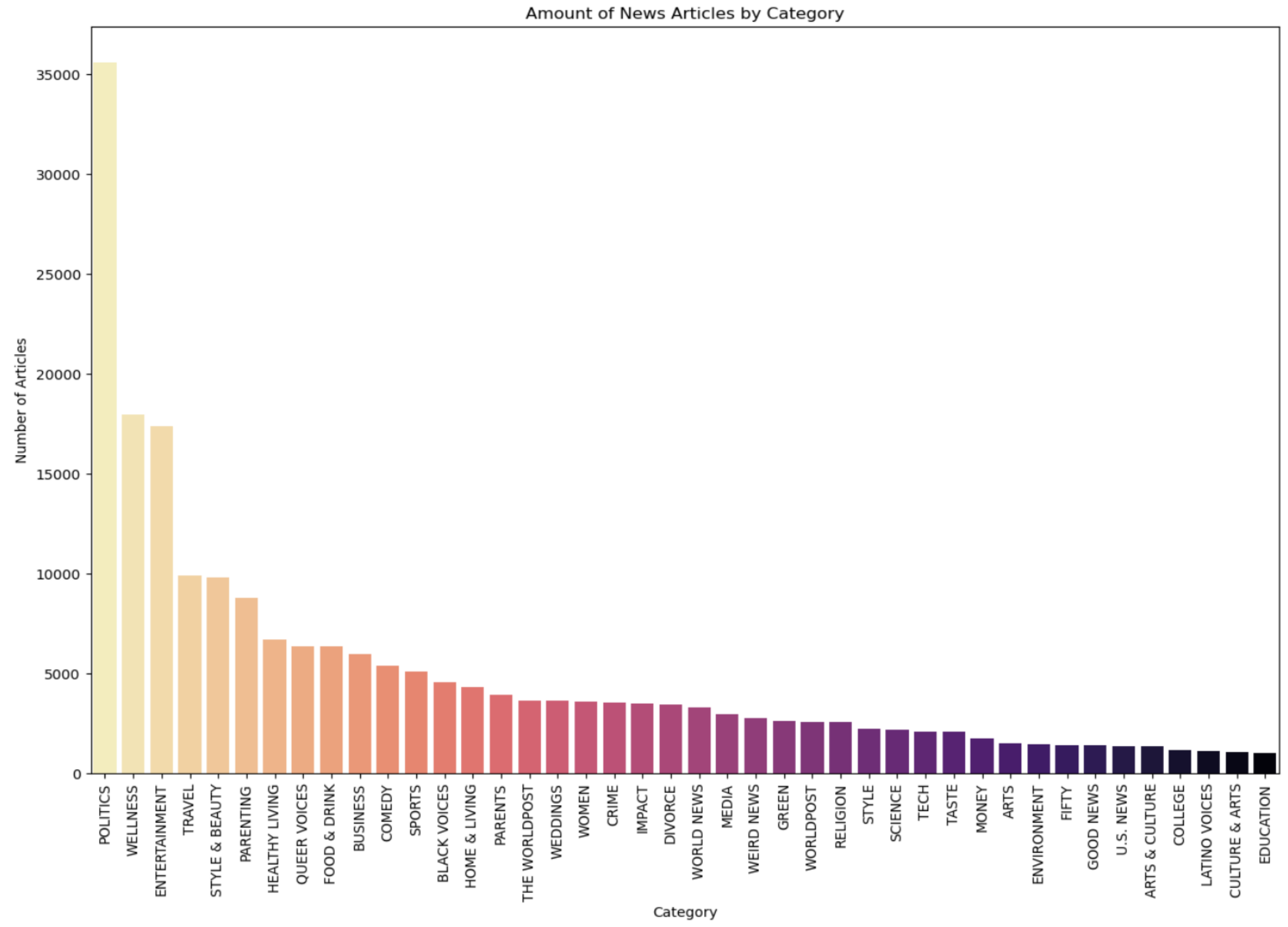


Figure 2 – The Number of News Articles per Class

- No missing values.
- 42 Categories. A large imbalance.
- Class imbalance techniques such as SMOTE will be investigated.
- Politics is by far the largest category.
- Many appear to be very similar such as "Arts & Culture" and "Culture & Arts".

BASELINE DATA PREPARATION

Text Cleaning and Preprocessing

This was broken down into several steps:

1. **Text Cleaning:** converts the text into lowercase, strips and removes punctuation.

2. **Expand contractions**: "I'd" --> "I would".

3. **Tokenization**: This separates words. These are called tokens.

4. **Stop word removal:** very common words that convey no meaning are removed e.g., “the”, "he", “you" or "on".

5. **Stemming:** This reduced words to their root form i.e., "shows", "showing" and "showed" will be reduced to "show".

6. **Lemmatization:** This also reduced words to their root form i.e., "better" and "best" will be reduced to "good".

7. **Bigrams and Trigrams:** Many sequential words should be treated as one such as "New" followed by "York" really means "New York".

8. **Removal of unique words:** words that occur only once in the entire dataset will be removed. There were approximately 25,000 unique words out of 63,000 words in total.

BASELINE DATA PREPARATION

Text Vectorization:

This is an important step to convert the words into numerical vectors for the machine learning process.

Three methods were tried in the initial phase:

- **Bag of Words:** A basic method.
- **Term Frequency-Inverse Document Frequency:** (tf-Idf). More advanced than BoW. It also adjusts word weights across the articles.
- **Word2vec:** A more advanced neural network based word embedding algorithm. This did not work too well on the baseline models.

BASELINE MODELS

Model	Classifier	Lemmatization Stemmming or Both	Contraction of Words	Remove Unique Words	Bigrams	Trigrams	Tfidf Vecotrizer or Other	Accuracy (%)	Comments
1	Logistic	Lemm	N	N	N	N	max_features=5000	58.30	
2	Multinomial NB	Lemm	N	N	N	N	max_features=5000	52.33	
3	Logistic	Lemm	Y	Y	Y	Y	max_features=5000	54.09	
4	Multinomial NB	Lemm	Y	Y	Y	Y	max_features=5000	48.38	
5	Logistic	Lemm	Y	N	N	N	max_features=5000	58.28	
6	Multinomial NB	Lemm	Y	N	N	N	max_features=5000	52.20	
7	Logistic	Lemm	N	Y	N	N	max_features=5000	58.32	
8	Multinomial NB	Lemm	N	Y	N	N	max_features=5000	52.33	
9	Logistic	Lemm	N	Y	Y	N	max_features=5000	55.30	
10	Multinomial NB	Lemm	N	Y	Y	N	max_features=5000	49.44	
11	Logistic	Lemm	N	Y	N	N	max_df=0.95, min_df=2	60.34	
12	Multinomial NB	Lemm	N	Y	N	N	max_df=0.95, min_df=2	44.56	
13	Logistic	Stem	N	Y	N	N	max_features=5000	58.83	
14	Multinomial NB	Stem	N	Y	N	N	max_features=5000	52.48	
15	Logistic	Stem	N	Y	N	N	max_features=20000	60.44	
16	Multinomial NB	Stem	N	Y	N	N	max_features=20000	47.65	
17	Logistic	Stem	N	Y	N	N	max_features=50000	60.51	This is the best Logistic Configuration
18	Multinomial NB	Stem	N	Y	N	N	max_features=50000	43.64	
19	Logistic	Stem	N	Y	N	N	max_df=0.95, min_df=2	60.38	
20	Multinomial NB	Stem	N	Y	N	N	max_df=0.95, min_df=2	45.06	
21	Logistic	Both	N	Y	N	N	max_features=5000	58.83	
22	Multinomial NB	Both	N	Y	N	N	max_features=5000	52.42	
23	Logistic	Both	N	Y	N	N	ngram_range=(1, 2)	NA	MUCH SLOWER!
24	Multinomial NB	Both	N	Y	N	N	ngram_range=(1, 2)	39.02	
25	Logistic	Both	N	Y	N	N	word2vec	52.64	
26	Multinomial NB	Both	N	Y	N	N	word2vec	NA	Cannot take negative values
27	Logistic	Stem	N	Y	N	N	bag of words	59.26	
28	Multinomial NB	Stem	N	Y	N	N	bag of words	57.81	This is the best NB configuration

Table 1 – Evaluation of Preprocessing and Vectorization Steps Applied to the Two Baseline Models

BASELINE RESULTS

Naïve-Bayes and Logistic Regression Models were used.

A good baseline was established.

The Logistic Regressor is clearly a better model for this dataset.

The best accuracy was 60.51% utilizing:

- Stemming
- Disabling Contraction of words
- Removal of unique words
- Bigrams and trigrams disabled. Enabling them resulted in a 4% degradation in both models.
- A Tf-idf vectorizer with max_features set to 50,000.

Class Imbalance affects the minority classes

MERGING OF SIMILAR CATEGORIES

- There are 42 categories, many of them are very similar.

```
['U.S. NEWS', 'COMEDY', 'PARENTING', 'WORLD NEWS', 'CULTURE & ARTS',
 'TECH', 'SPORTS', 'ENTERTAINMENT', 'POLITICS', 'WEIRD NEWS',
 'ENVIRONMENT', 'EDUCATION', 'CRIME', 'SCIENCE', 'WELLNESS',
 'BUSINESS', 'STYLE & BEAUTY', 'FOOD & DRINK', 'MEDIA',
 'QUEER VOICES', 'HOME & LIVING', 'WOMEN', 'BLACK VOICES', 'TRAVEL',
 'MONEY', 'RELIGION', 'LATINO VOICES', 'IMPACT', 'WEDDINGS',
 'COLLEGE', 'PARENTS', 'ARTS & CULTURE', 'STYLE', 'GREEN', 'TASTE',
 'HEALTHY LIVING', 'THE WORLDPOST', 'GOOD NEWS', 'WORLDPOST',
 'FIFTY', 'ARTS', 'DIVORCE'], dtype=object)
```

```
'PARENTS': 'PARENTING',
'THE WORLDPOST': 'WORLD NEWS',
'WORLDPOST': 'WORLD NEWS',
'BUSINESS': 'BUSINESS & FINANCE',
'MONEY': 'BUSINESS & FINANCE',
'COLLEGE': 'EDUCATION',
'STYLE': 'STYLE & BEAUTY',
'GREEN': 'ENVIRONMENT',
'ARTS': 'ARTS & CULTURE',
'CULTURE & ARTS': 'ARTS & CULTURE',
'HEALTHY LIVING': 'WELLNESS',
'TASTE': 'FOOD & DRINK'
```

- There were reduced down to 31 categories.
- Overall Accuracy improved from 60.51% to 66.76%.
- Similar gains in Precision, Recall and F1-Score.

ADDRESSING THE CLASS IMBALANCE

- There is still a large class imbalance.

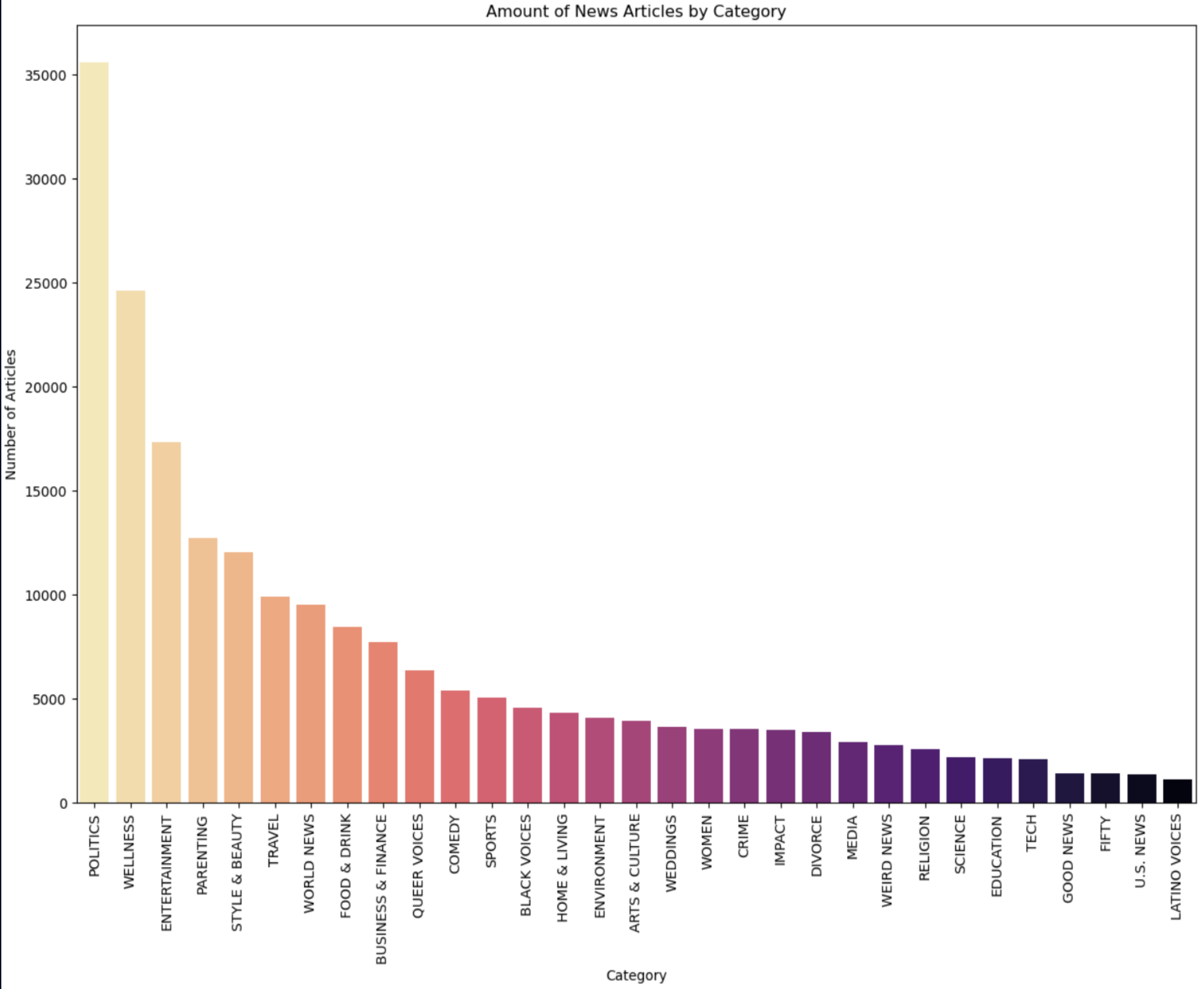


Figure 3 – The Number of News Articles per Class after Merging

BUSINESS & FINANCE	1656
COMEDY	1656
WEDDINGS	1656
IMPACT	1656
RELIGION	1656
TRAVEL	1656
QUEER VOICES	1656
WELLNESS	1656
WOMEN	1656
HOME & LIVING	1656
BLACK VOICES	1656
MEDIA	1656
STYLE & BEAUTY	1656
WORLD NEWS	1656
DIVORCE	1656
FOOD & DRINK	1656
CRIME	1656
PARENTING	1656
SCIENCE	1656
EDUCATION	1656
ENTERTAINMENT	1656
WEIRD NEWS	1656
ENVIRONMENT	1656
SPORTS	1656
POLITICS	1656
ARTS & CULTURE	1656
TECH	1656
GOOD NEWS	1397
FIFTY	1396
U.S. NEWS	1377
LATINO VOICES	1130

- SMOTE or oversampling is not an option due to much larger dataset size.
- Undersampling was tried.
- Reduction of dataset to 25% of original size
- 10% degradation in accuracy.
- Far fewer words in the training.
- Recall did rise slightly.
- This was not implemented.

DEEP LEARNING

Encoding, Tokenization and Sequence Padding

- Labels encoded as integers
- Unique words tokenized as integers
- Padding ensure all text entries are of equal length

Word Embedding

- Words are represented as vectors.
- Semantic and syntactic similarity are identified between words.
- Relationships with other words detected.
- Word2Vec
- GloVe – Global Vectors for word representation

Neural Network Models

Keras package was used for RNNs

These models are designed for sequences.

- BERT – pre-trained model. Far too slow!
- GRU – Gated Recurrent Unit.
- Long Short-Term Memory – more complex than GRU.

DEEP LEARNING RESULTS

Model	Classifier	Neurons	Tokenizer / Word Embedder	num_words	oov_token	Max Length	learning rate	embedding dimensions	Optimizer	dropout / recurrent dropout	Test Accuracy (%)	Comments
23	GRU Bi-directional	32	Tokenizer	all words	<UNK>	90th percentile	0.001	300	Adam	0.2	67.70	slower than 100
24	GRU Bi-directional	32	Tokenizer	all words	<UNK>	90th percentile	0.001	300	Adam	0	67.76	slower than 100
25	GRU Bi-directional	32	Tokenizer	all words	<UNK>	90th percentile	0.001	300	RMSProp	0	67.85	slower than 100
26	GRU Bi-directional	32	Tokenizer	all words	<UNK>	90th percentile	0.001	500	RMSProp	0	68.05	much slower than 100
27	GRU Bi-directional	32	Tokenizer	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	68.37	much slower than 500
28	GRU Bi-directional	32	Tokenizer	all words	<UNK>	90th percentile	0.001	3000	RMSProp	0	68.36	much slower than 1000, 4 mins per epoch to over 10 mins
29	GRU Bi-directional	64	Tokenizer	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	68.75	much slower than 500
30	GRU Bi-directional	128	Tokenizer	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	69.16	much slower than 500
31	GRU Bi-directional	32	Tokenizer /GloVe	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	62.77	GloVe decreases performance
32	LSTM - Bi-directional	128/64	Tokenizer	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	67.68	GloVe decreases performance
33	LSTM - Bi-directional	128	Tokenizer /GloVe	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	63.29	much slower than 500
34	GRU Bi-directional	128/64	Tokenizer /GloVe	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	68.03	much slower than 500
35	GRU Bi-directional	256	Tokenizer /GloVe	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	68.88	much slower than 500
36	GRU Bi-directional	128	Tokenizer /GloVe	all words	<UNK>	90th percentile	0.001	1000	Adam	0	68.29	not as good as RMSProp
37	LSTM - Bi-directional	128	Tokenizer	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	68.64	The best LSTM, but GRU is better
38	GRU Bi-directional	128	Word2Vec	all words	<UNK>	90th percentile	0.001	1000	RMSProp	0	65.18	fast but less accurate
39	GRU Bi-directional	128	Tokenizer	all words	<UNK>	99th percentile	0.001	1000	RMSProp	0	69.31	much slower than 500
40	GRU Bi-directional	128	Tokenizer	all words	<UNK>	max	0.001	1000	RMSProp	0	69.32	much slower than 99th percentile
41	LSTM Bi-directional	128	Tokenizer	all words	<UNK>	max	0.001	1000	RMSProp	0	69.16	much slower than 99th percentile

Over 40 combinations of Word Embedders, model architectures and hyperparameters were tried.

Convergence was very fast. Usually after only 2 epochs.

Keras ModelCheckpoint function, saves the best configuration.

Table 2 – Evaluation of RNNs, Hyperparameters and Word Embedding Techniques

DEEP LEARNING RESULTS

GloVe and Word2Vec did not work well. Typically 4 - 5% worse.

GRU was consistently better than the LSTM. Typically by 1%.

Best results with:

- GRU – Bi-directional.
- 1 layer architecture, 128 units.
- RMSProp solver.
- Embedding dimension of 1000 words.
- Max length of all sequences worked best. No truncations.

Best Test Set Accuracy of 69.32%

This is better than the Logistic Regression. 66.76%

PREDICTION EXAMPLES

processed_combined_info	true_labels	predicted_labels
elder independ establish common ground whether care age parent neighbor apart build help home health aid connect ill spous look establish common ground drive forward success caregiv relationship	WELLNESS	WELLNESS
florida woman bitten shark inner tube	CRIME	WEIRD NEWS
sandal revolut may look go heck thing precis fashion commun put head togeth give shoe name reader like present newest hybrid shoe luxuri flatform	STYLE & BEAUTY	STYLE & BEAUTY
13 reason goe beyond tape haunt season 2 trailer new season debut netflix may 18	ENTERTAINMENT	ENTERTAINMENT
fowl bizarr bowl footbal next big sport mich ap detroit area entrepreneur believ score touchdown new busi idea thrown	WEIRD NEWS	SPORTS
st regi princevil pauper vill st regi love horizont ivori tower isol tip princevil leav kauai alway want	TRAVEL	TRAVEL
nobel prize medicin jame rothman randi thoma jointli win prize nobel committe said research traffic transport system cell help scientist understand	WELLNESS	SCIENCE
start viral pay homag youtub meme video parodi billi joel start fire go stuck head day	COMEDY	COMEDY
poison daughter russian spi releas hospit end treatment mark signific mileston	WORLD NEWS	ENTERTAINMENT
ed sheeran grammi nomin complet surpris expect year alreadi drunk found didnt expect year sheeran told huffington post z100	ENTERTAINMENT	ENTERTAINMENT
challeng patel poll show tighten race key new york primari patel take rep carolyn maloney jerri nadler democrat nomin manhattan congression district	POLITICS	POLITICS
trayvon martin case mean middl class black cri heard zimmerman verdict knew would tear countri apart racial line also knew middl class black person go call upon perspect white colleagu friend famili	BLACK VOICES	BLACK VOICES

Stemming used

Mispredicted labels:

2, 5, 7, 9

Very difficult to classify some of these descriptions accurately

Training on the full article would likely yield better results.

CONCLUSION

Data Preprocessing and Feature Engineering

Merging of Categories: provided the largest improvement. Highlights the importance of good data preparation.

Modelling

Logistic Regression: Compared very well to GRU and LSTM

Recurrent Neural Networks: GRU than LSTM worked better for smaller text sequences.

Hyperparameter Tuning: Improvements were definitely gained.

Improvement from baseline: 60.51% to 69.32%

END OF THE PRESENTATION

THANK YOU