Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Implementation and Evaluation of Text Classifiers

NLP Project

02.07.2020

Amirhossein Pakdaman

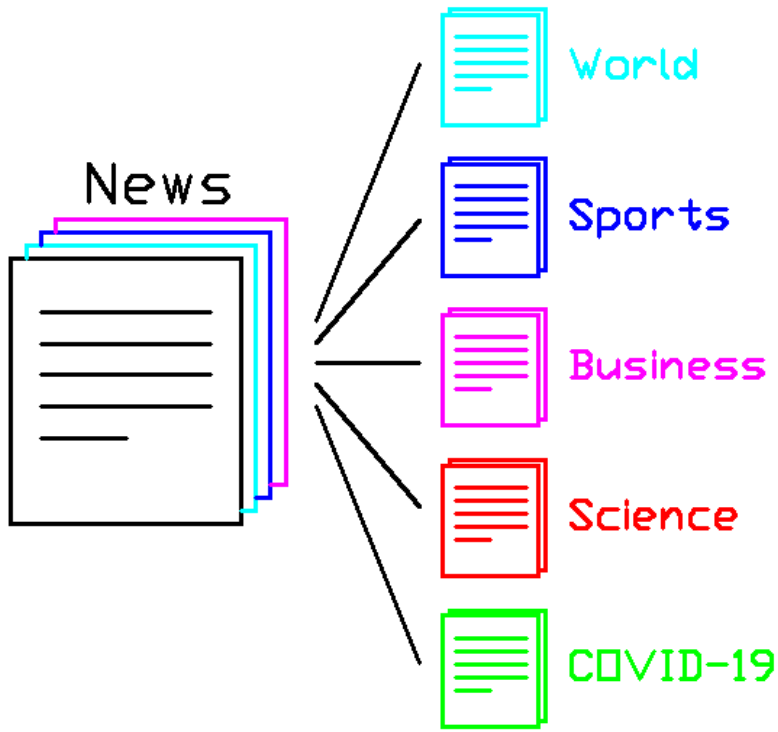
Introduction

Target:

- Implementing a **news classifier** with different techniques.
- Compare and benchmarking the results and performance of each method.

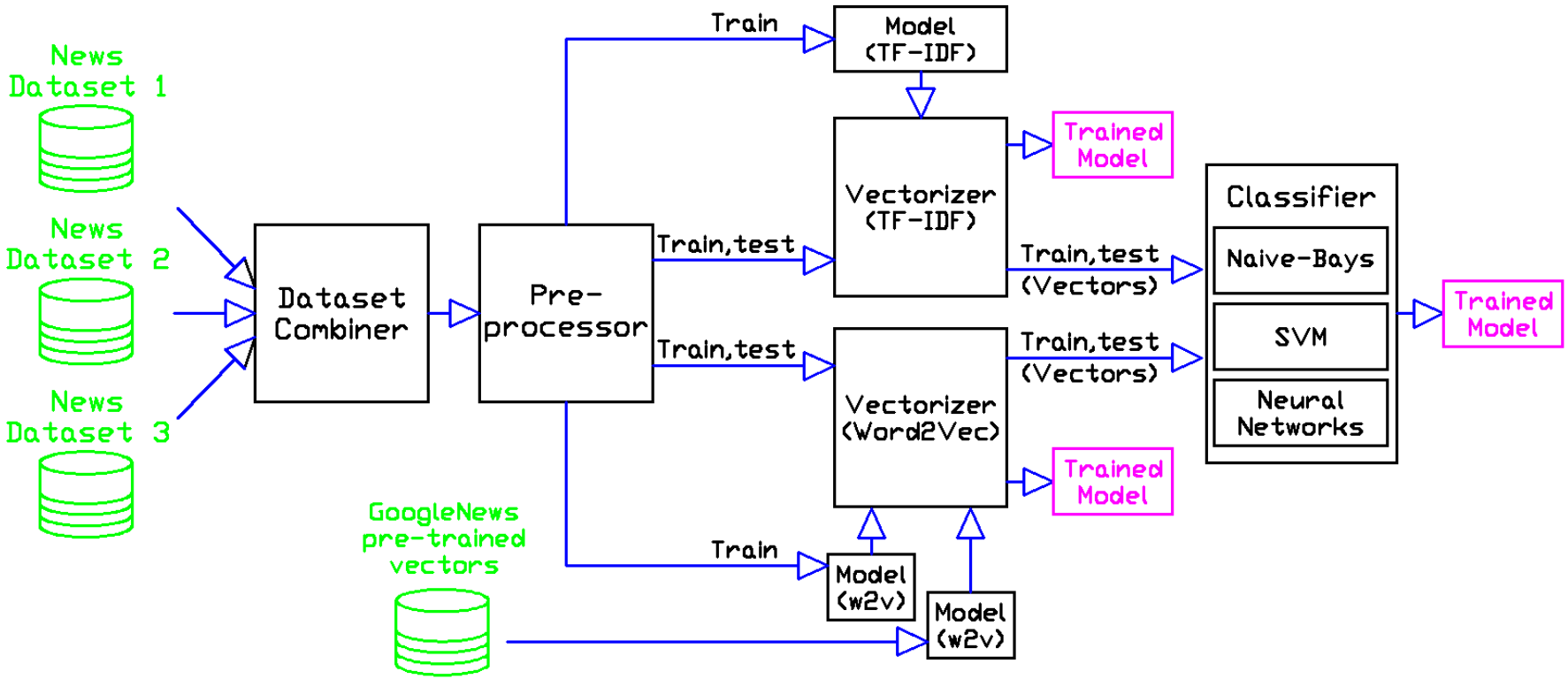
Classes:

- World
- Sports
- Business
- Science
- COVID-19



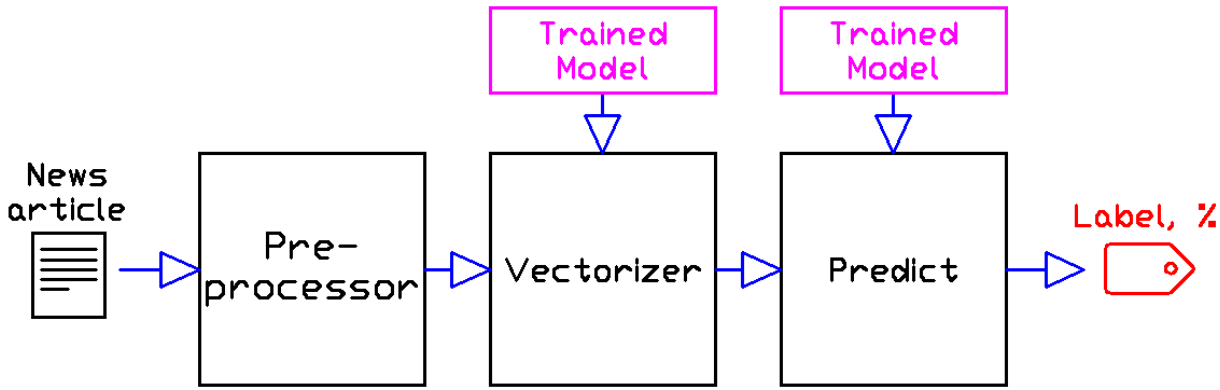
Pipeline

Training phase:



Pipeline

Inference phase:



Datasets

Input datasets:

- AG's News Topic Classification Dataset [1]
  - Consists of 4 classes: World, Sport, Business, Science.
  - Each class contains 30,000 items.
  - Collected in 2015.
- COVID-19 News Articles Open Research Dataset [2]
  - Contains over 7,000 news articles about COVID-19.
  - Collected from CBC news.
- COVID-19 Public Media Dataset by Anacode [3]
  - Contains articles about COVID-19.
  - News articles from BBC, CNN, CNBC, and Reuters are picked.

Final dataset:

- 5 classes, each contain 30,000 items.
- Due to memory limitations, a reduced version is used for training:
  - Train set: 50,000
  - Test set: 5,000

[1]: https://www.kaggle.com/amananandrai/ag-news-classification-dataset
[2]: https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26
[3]: https://www.kaggle.com/jannalipenkova/covid19-public-media-dataset

Pre-process

0. Raw text:

```
Dolphin groups, or "pods", rely on 20 sociali ...
Tyrannosaurus rex achieved its massive size du...
```

1. Tokenization:

```
[Dolphin, groups„ or, "pods"„ rely, on, 20, ...
[Tyrannosaurus, rex, achieved, its, massive, s...
```

2. To lower-case:

```
[dolphin, groups„ or, "pods"„ rely, on, 20, ...
[tyrannosaurus, rex, achieved, its, massive, s...
```

3. Punctuation removal:

```
[dolphin, groups, or, pods, rely, on, 20, lit...
[tyrannosaurus, rex, achieved, its, massive, s...
```

4. Number digits removal:

```
[dolphin, groups, or, pods, rely, on, socialit...
[tyrannosaurus, rex, achieved, its, massive, s...
```

5. Lemmatization:

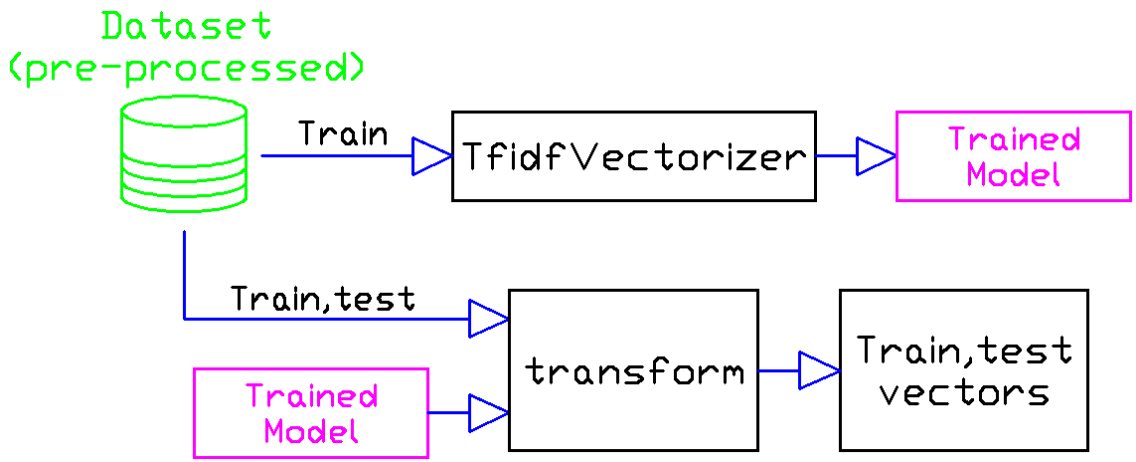
```
[dolphin, group, or, pod, rely, on, socialite,...
[tyrannosaurus, rex, achieved, it, massive, si...
```

6. Stop words removal:

```
[dolphin, group, pod, rely, socialite, keep, c...
[tyrannosaurus, rex, achieved, massive, size, ...
```

TF-IDF Vectorization

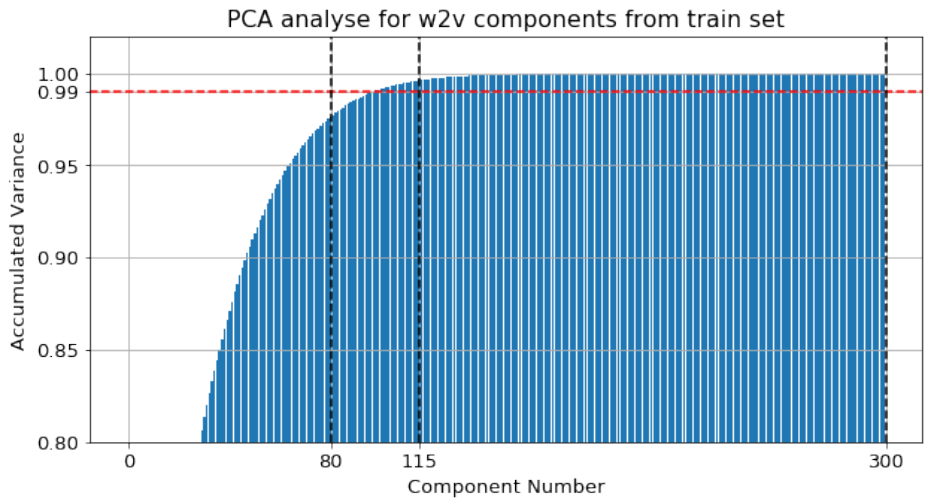
- By **TfidfVectorizer** from Scikit-Learn.
- Model is learned from **train set**.
- Three different feature lengths are generated: 300, 1000, 8000.
- The whole matrix of vectors is stored on memory.
  - May cause memory error.



Word2Vec Vectorization

Training Word2Vec model:

1. Pre-trained vectors from Google. [1]
   - Contains 300-dimensional vectors for 3 million words and phrases.
2. Train set data.
   - Causes a dependency between vectorizer and classifier.
   - Three different feature lengths: 80, 115, 300



Word averaging method:

- Converts word embeddings to get a vector per example.
- Takes average vector of all tokens in one example.

[1]: https://code.google.com/archive/p/word2vec

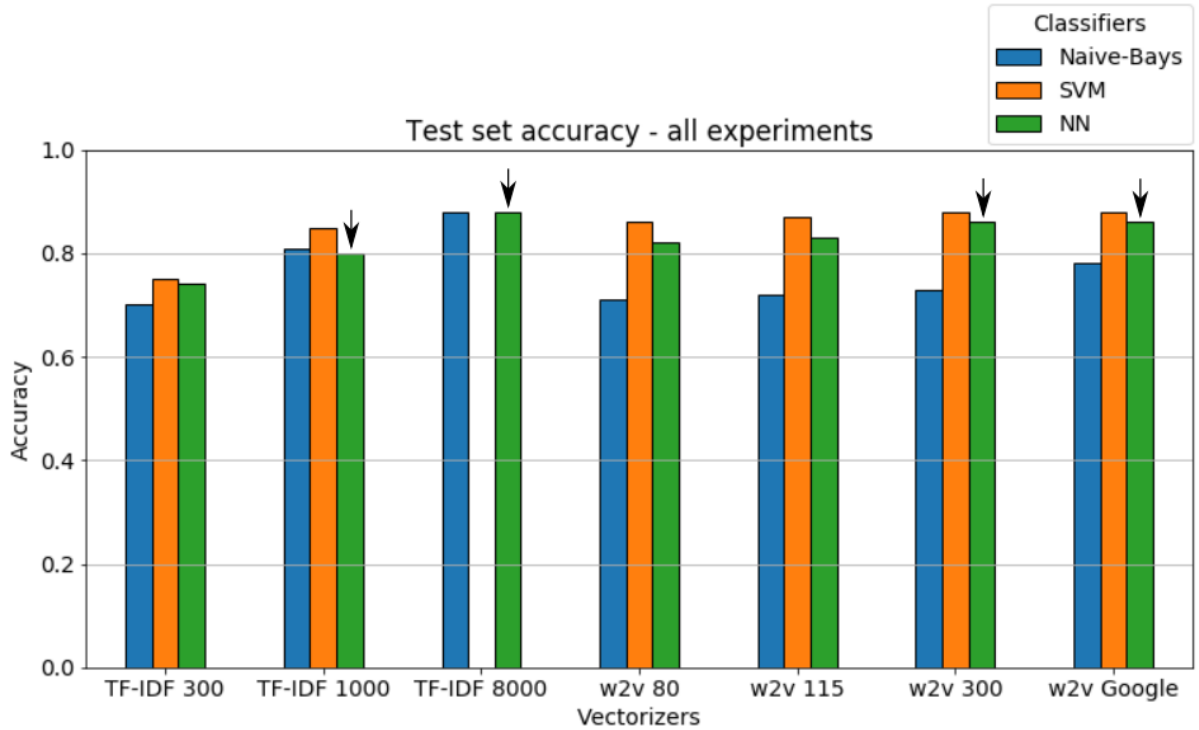
Vectorization - Summary

- Vectorization is done in 3 different ways, and with different vector lengths.
- Resulting in 7 different vectorized datasets.
- Each will be fed to the classifiers individually.

Model	Dim.	Size	Train time
TF-IDF	300	1535KB	4.8s
	1000	1568KB	4.9s
	8000	1897KB	5.3s
w2v, Train set	80	25.8MB	28s
	115	36MB	35s
	300	90MB	64s
w2v, Google	300	533MB	34s (load time)

Classification

- Three classification methods: Naive-Bays, SVM, NN
- Applying on all vectorized data, resulting 21 classifiers.



Classification - Training

Naive-Bays

Vectorizer	Dim.	Train time	Predict time	Memory usage	Test accuracy
TF-IDF	300	258ms	11ms	25KB	0.70
	1000	289ms	31ms	80KB	0.81
	8000	3.5s	167ms	641KB	0.88
w2v, Train set	80	70ms	10ms	7KB	0.71
	115	72ms	10ms	10KB	0.72
	300	164ms	19ms	25KB	0.73
w2v, Google	300	218ms	17ms	25KB	0.78

SVM

Vectorizer	Dim.	Train time	Predict time	Memory usage	Test accuracy
TF-IDF	300	37m	1m 30s	74MB	0.75
	1000	3h 9m	6m 50s	235MB	0.85
	8000	-	-	-	-
w2v, Train set	80	3m 46s	18s	12.6MB	0.86
	115	4m 44s	24s	17.3MB	0.87
	300	9m 12s	57s	41.3MB	0.88
w2v, Google	300	13m 39s	57s	44.5MB	0.88

NN

Vectorizer	Dim.	Train time	Predict time	Memory usage	Test accuracy
TF-IDF	300	8m 19s	167ms	138KB	0.74
	1000	9m 25s	273ms	406KB	**0.80**
	8000	9m 35s	571ms	3094KB	**0.88**
w2v, Train set	80	8m 1s	243ms	53KB	0.82
	115	7m 48s	189ms	67KB	0.83
	300	7m 50s	187ms	138KB	**0.86**
w2v, Google	300	5m 3s	140ms	138KB	**0.86**

Classification - Inference

- The inference phase is examined with 4 best classifiers.
- Recent articles from DW news website are used.
- For each class 5 articles are tested.

Classifier (NN)	Load	Inference	Inference accuracy						Test acc.
			World	Sports	Business	Science	COVID-19	Avg.	
TF-IDF 1000	8s	25ms	3/5	5/5	1/5	3/5	5/5	0.68	0.80
TF-IDF 8000	8.3s	25ms	5/5	5/5	4/5	4/5	5/5	0.92	0.88
w2v 300	**4.8s**	**18ms**	**1/5**	**0/5**	**2/5**	**4/5**	**1/5**	**0.32**	**0.86**
w2v Google	42.3s	20ms	5/5	4/5	3/5	4/5	0/5	0.64	0.86

Analysis:

- **w2v 300:** The dependency between vectorizer and classifier:
  - Low accuracy for an unseen domain.
  - High test accuracy is because of familiar domain.
- **w2v Google:** COVID19 related words are not included in the pre-trained dataset.
- **TF-IDF 8000:** Gives very good results with unseen domain.

Classification - Inference

- The inference phase is examined with 4 best classifiers.
- Recent articles from DW news website are used.
- For each class 5 articles are tested.

Classifier (NN)	Load	Inference	Inference accuracy						Test acc.
			World	Sports	Business	Science	COVID-19	Avg.	
TF-IDF 1000	8s	25ms	3/5	5/5	1/5	3/5	5/5	0.68	0.80
TF-IDF 8000	8.3s	25ms	5/5	5/5	4/5	4/5	5/5	0.92	0.88
w2v 300	4.8s	18ms	1/5	0/5	2/5	4/5	1/5	0.32	0.86
w2v Google	**42.3s**	**20ms**	**5/5**	**4/5**	**3/5**	**4/5**	**0/5**	**0.64**	**0.86**

Analysis:

- **w2v 300:** The dependency between vectorizer and classifier:
  - Low accuracy for an unseen domain.
  - High test accuracy is because of familiar domain.
- **w2v Google:** COVID19 related words are not included in the pre-trained dataset.
- **TF-IDF 8000:** Gives very good results with unseen domain.

Classification - Inference

- The inference phase is examined with 4 best classifiers.
- Recent articles from DW news website are used.
- For each class 5 articles are tested.

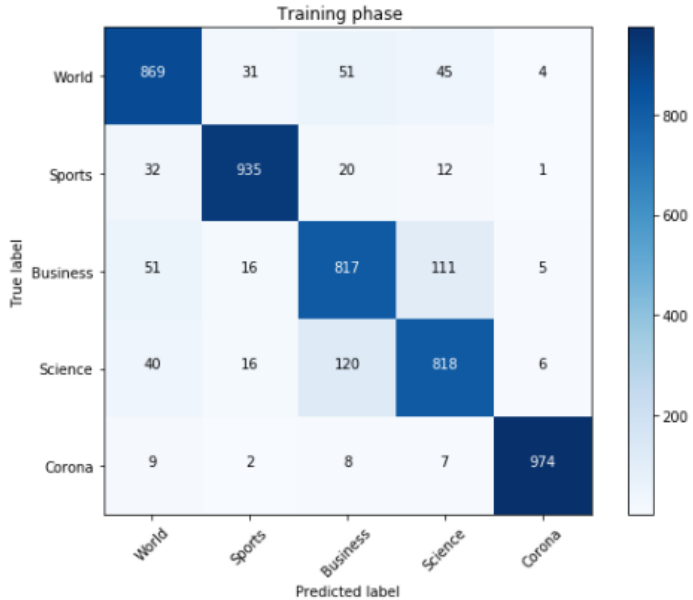
Classifier (NN)	Load	Inference	Inference accuracy						Test acc.
			World	Sports	Business	Science	COVID-19	Avg.	
TF-IDF 1000	8s	25ms	3/5	5/5	1/5	3/5	5/5	0.68	0.80
TF-IDF 8000	**8.3s**	**25ms**	**5/5**	**5/5**	**4/5**	**4/5**	**5/5**	**0.92**	**0.88**
w2v 300	4.8s	18ms	1/5	0/5	2/5	4/5	1/5	0.32	0.86
w2v Google	42.3s	20ms	5/5	4/5	3/5	4/5	0/5	0.64	0.86

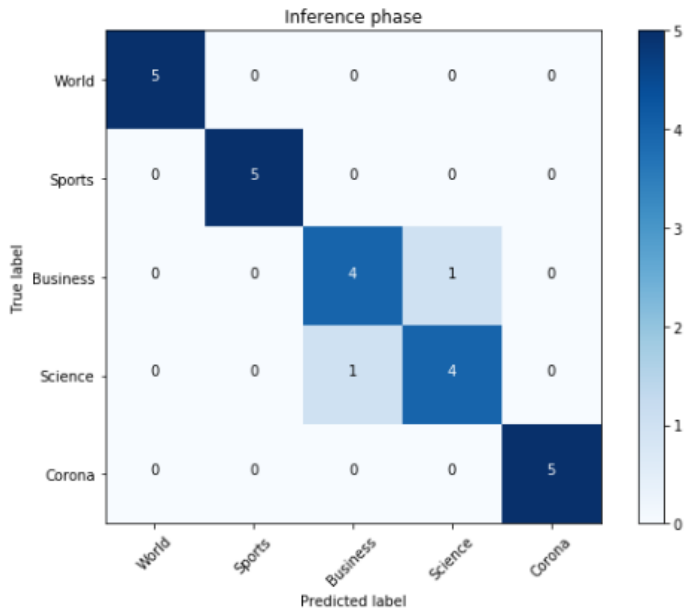
Analysis:

- **w2v 300:** The dependency between vectorizer and classifier:
  - Low accuracy for an unseen domain.
  - High test accuracy is because of familiar domain.
- **w2v Google:** COVID19 related words are not included in the pre-trained dataset.
- **TF-IDF 8000:** Gives very good results with unseen domain.

Best Result

- Vectorization:
  - Type: TF-IDF
  - Vectors length: 8000
  - Memory usage: 1897KB
  - Train time: 5.3s
- Classification, training phase:
  - Type: Neural networks
  - Test accuracy: 0.88
  - Memory usage: 3094KB
  - Train time: 9m 35s
- Classification, inference phase:
  - Data: 25 news articles
  - Load time: 8.3s
  - Inference time: 25ms
  - Accuracy: 0.92





Conclusions

- Words and phrases in news articles change over time.
- Word2Vec models are learned by the content of the corpus.
- Training a Word2Vec model on domain-specific data causes dependencies.
- Extracting sentence embedding by word averaging results in loosing information.
- TF-IDF vectors are not dependent on the whole content.
- TF-IDF vectorization gives better results in case of news classification.
- A vectorized dataset is stored on memory which can cause memory error.

References

GitHub repository of this project:
https://github.com/amirhpd/text_classifier

Datasets:
https://www.kaggle.com/amananandrai/ag-news-classification-dataset
https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26
https://www.kaggle.com/jannalipenkova/covid19-public-media-dataset
https://code.google.com/archive/p/word2vec

Guidance from:
https://www.kaggle.com/amananandrai/news-article-classifier-with-different-models
https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568