Curation
CORPORATION
feed your competitive edge

Business News Classification Engine

Springboard Capstone Project 2

James Flint | mail@jamesflint.Net | 2018-03-30

REQUIREMENT

- Take CurationCorp’s labelled news database of 43,502 articles
- Train a neural net-based topic classification engine
- Make this functionality available via a cloud-based API

APPROACH

- **Data wrangling**
- **Compare Classifiers**
    - A multi-layer neural net (NN)
    - A convolutional neural net (CNN)
    - A long/short term memory neural net (LSTM)
    - A very deep convolutional neural net (VDCNN)
- **Build a prediction API**
- **Future research**

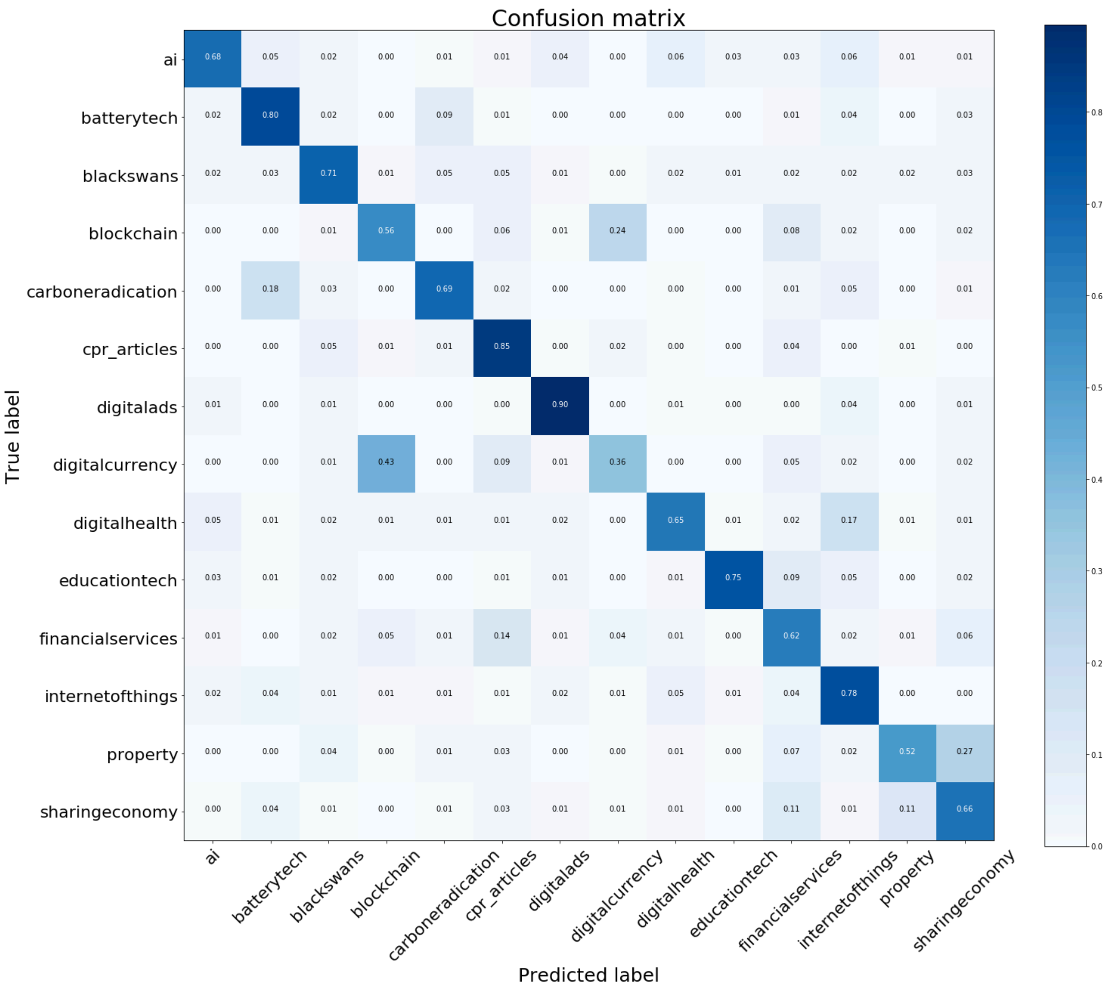
UNBALANCED DATASET

cpr_articles	6846
blackswans	4837
batterytech	4092
financialservices	3986
carboneradication	3690
sharingeconomy	3574
digitalads	2920
internetofthings	2627
property	2132
digitalhealth	1943
digitalcurrency	1914
ai	1722
blockchain	1650
educationtech	1555

BALANCED DATASET

batterytech	16368
financialservices	15944
internetofthings	15762
digitalhealth	15544
digitalcurrency	15312
property	14924
carboneradication	14760
digitalads	14600
blackswans	14511
sharingeconomy	14296
educationtech	13995
ai	13776
cpr_articles	13692
blockchain	13200

MULTILAYER NEURAL NET



	ai	batterytech	blackswans	blockchain	carboneradication	cpr_articles	digitalads	digitalcurrency	digitalhealth	educationtech	financialservices	internetofthings	property	sharingeconomy
ai	0.68	0.05	0.02	0.00	0.01	0.01	0.04	0.00	0.06	0.03	0.03	0.06	0.01	0.01
batterytech	0.02	0.80	0.02	0.00	0.09	0.01	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.03
blackswans	0.02	0.03	0.71	0.01	0.05	0.05	0.01	0.00	0.02	0.01	0.02	0.02	0.02	0.03
blockchain	0.00	0.00	0.01	0.56	0.00	0.06	0.01	0.24	0.00	0.00	0.08	0.02	0.00	0.02
carboneradication	0.00	0.18	0.03	0.00	0.69	0.02	0.00	0.00	0.00	0.00	0.01	0.05	0.00	0.01
cpr_articles	0.00	0.00	0.05	0.01	0.01	0.85	0.00	0.02	0.00	0.00	0.04	0.00	0.01	0.00
digitalads	0.01	0.00	0.01	0.00	0.00	0.00	0.90	0.00	0.01	0.00	0.00	0.04	0.00	0.01
digitalcurrency	0.00	0.00	0.01	0.43	0.00	0.09	0.01	0.36	0.00	0.00	0.05	0.02	0.00	0.02
digitalhealth	0.05	0.01	0.02	0.01	0.01	0.01	0.02	0.00	0.65	0.01	0.02	0.17	0.01	0.01
educationtech	0.03	0.01	0.02	0.00	0.00	0.01	0.01	0.00	0.01	0.75	0.09	0.05	0.00	0.02
financialservices	0.01	0.00	0.02	0.05	0.01	0.14	0.01	0.04	0.01	0.00	0.62	0.02	0.01	0.06
internetofthings	0.02	0.04	0.01	0.01	0.01	0.01	0.02	0.01	0.05	0.01	0.04	0.78	0.00	0.00
property	0.00	0.00	0.04	0.00	0.01	0.03	0.00	0.00	0.01	0.00	0.07	0.02	0.52	0.27
sharingeconomy	0.00	0.04	0.01	0.00	0.01	0.03	0.01	0.01	0.01	0.00	0.11	0.01	0.11	0.66

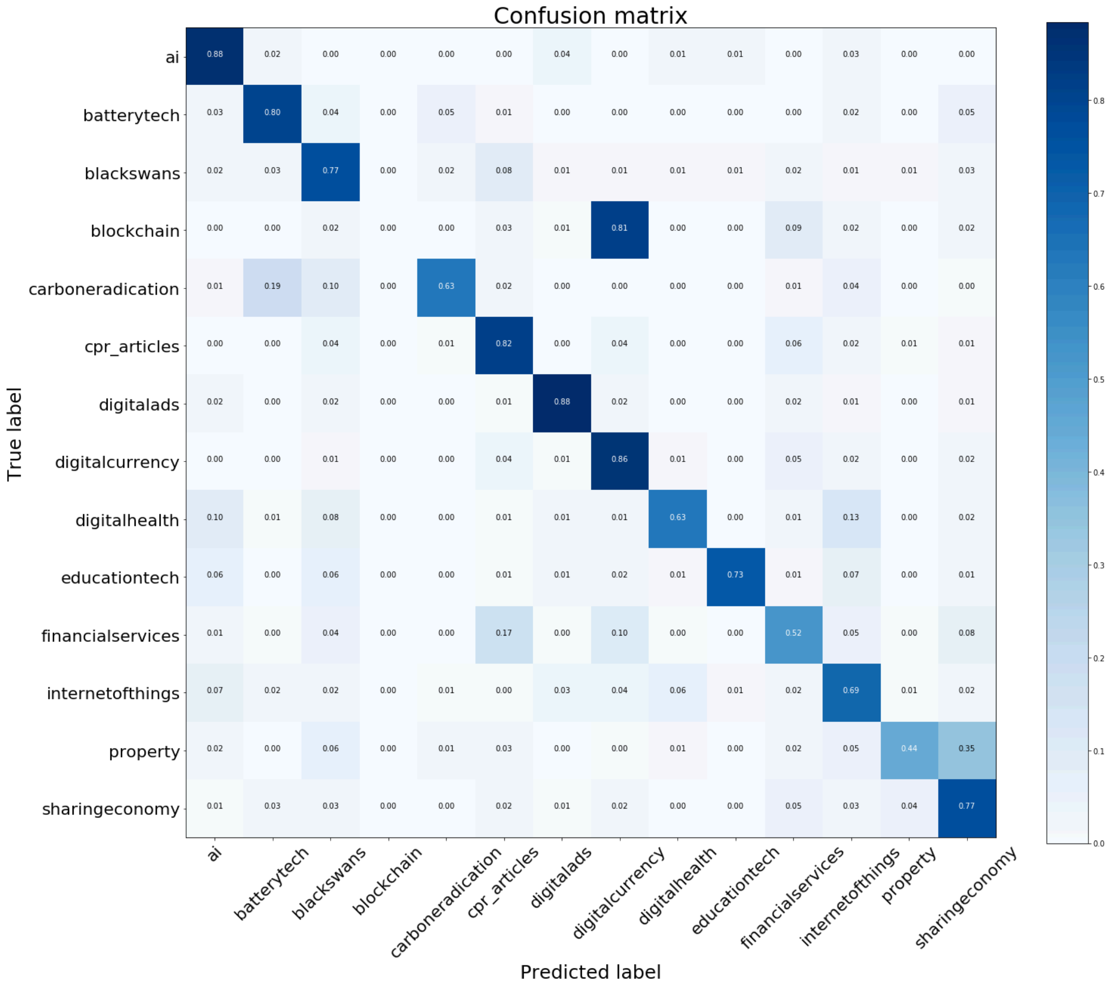
Best performance:

72.5% accuracy (unbalanced data)

Support vector machine benchmark:

77% accuracy

CONVOLUTIONAL NEURAL NET



	ai	batterytech	blackswans	blockchain	carboneradication	cpr_articles	digitalads	digitalcurrency	digitalhealth	educationtech	financialservices	internetofthings	property	sharingeconomy
ai	0.88	0.02	0.00	0.00	0.00	0.00	0.04	0.00	0.01	0.01	0.00	0.03	0.00	0.00
batterytech	0.03	0.80	0.04	0.00	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.05
blackswans	0.02	0.03	0.77	0.00	0.02	0.08	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.03
blockchain	0.00	0.00	0.02	0.00	0.00	0.03	0.01	0.81	0.00	0.00	0.09	0.02	0.00	0.02
carboneradication	0.01	0.19	0.10	0.00	0.63	0.02	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.00
cpr_articles	0.00	0.00	0.04	0.00	0.01	0.82	0.00	0.04	0.00	0.00	0.06	0.02	0.01	0.01
digitalads	0.02	0.00	0.02	0.00	0.00	0.01	0.88	0.02	0.00	0.00	0.02	0.01	0.00	0.01
digitalcurrency	0.00	0.00	0.01	0.00	0.00	0.04	0.01	0.86	0.01	0.00	0.05	0.02	0.00	0.02
digitalhealth	0.10	0.01	0.08	0.00	0.00	0.01	0.01	0.01	0.63	0.00	0.01	0.13	0.00	0.02
educationtech	0.06	0.00	0.06	0.00	0.00	0.01	0.01	0.02	0.01	0.73	0.01	0.07	0.00	0.01
financialservices	0.01	0.00	0.04	0.00	0.00	0.17	0.00	0.10	0.00	0.00	0.52	0.05	0.00	0.08
internetofthings	0.07	0.02	0.02	0.00	0.01	0.00	0.03	0.04	0.06	0.01	0.02	0.69	0.01	0.02
property	0.02	0.00	0.06	0.00	0.01	0.03	0.00	0.00	0.01	0.00	0.02	0.05	0.44	0.35
sharingeconomy	0.01	0.03	0.03	0.00	0.00	0.02	0.01	0.02	0.00	0.00	0.05	0.03	0.04	0.77

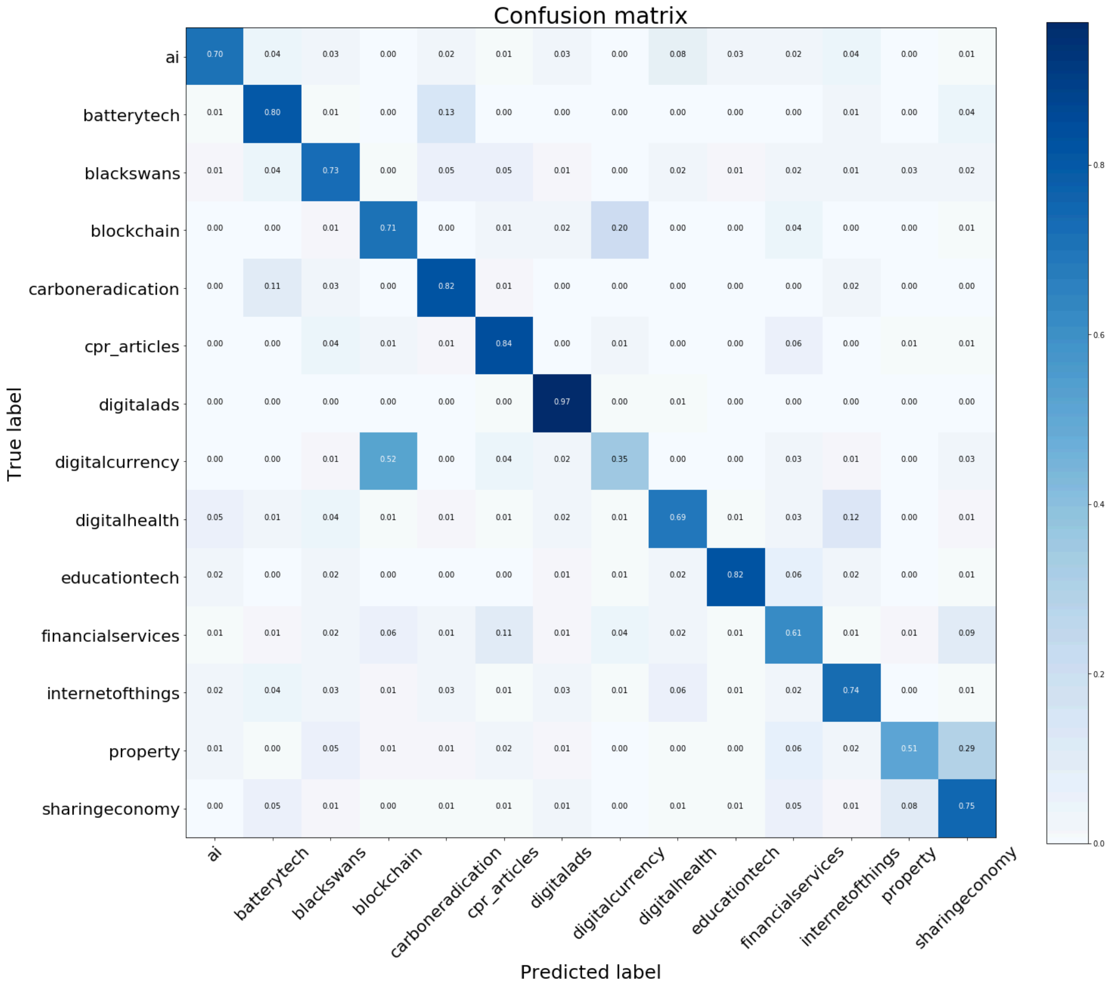
Best performance:

72.6% accuracy (unbalanced data)

Support vector machine benchmark:

77% accuracy

LONG SHORT-TERM MEMORY



	ai	batterytech	blackswans	blockchain	carboneradication	cpr_articles	digitalads	digitalcurrency	digitalhealth	educationtech	financialservices	internetofthings	property	sharingeconomy
ai	0.70	0.04	0.03	0.00	0.02	0.01	0.03	0.00	0.08	0.03	0.02	0.04	0.00	0.01
batterytech	0.01	0.80	0.01	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.04
blackswans	0.01	0.04	0.73	0.00	0.05	0.05	0.01	0.00	0.02	0.01	0.02	0.01	0.03	0.02
blockchain	0.00	0.00	0.01	0.71	0.00	0.01	0.02	0.20	0.00	0.00	0.04	0.00	0.00	0.01
carboneradication	0.00	0.11	0.03	0.00	0.82	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
cpr_articles	0.00	0.00	0.04	0.01	0.01	0.84	0.00	0.01	0.00	0.00	0.06	0.00	0.01	0.01
digitalads	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.01	0.00	0.00	0.00	0.00	0.00
digitalcurrency	0.00	0.00	0.01	0.52	0.00	0.04	0.02	0.35	0.00	0.00	0.03	0.01	0.00	0.03
digitalhealth	0.05	0.01	0.04	0.01	0.01	0.01	0.02	0.01	0.69	0.01	0.03	0.12	0.00	0.01
educationtech	0.02	0.00	0.02	0.00	0.00	0.00	0.01	0.01	0.02	0.82	0.06	0.02	0.00	0.01
financialservices	0.01	0.01	0.02	0.06	0.01	0.11	0.01	0.04	0.02	0.01	0.61	0.01	0.01	0.09
internetofthings	0.02	0.04	0.03	0.01	0.03	0.01	0.03	0.01	0.06	0.01	0.02	0.74	0.00	0.01
property	0.01	0.00	0.05	0.01	0.01	0.02	0.01	0.00	0.00	0.00	0.06	0.02	0.51	0.29
sharingeconomy	0.00	0.05	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.05	0.01	0.08	0.75

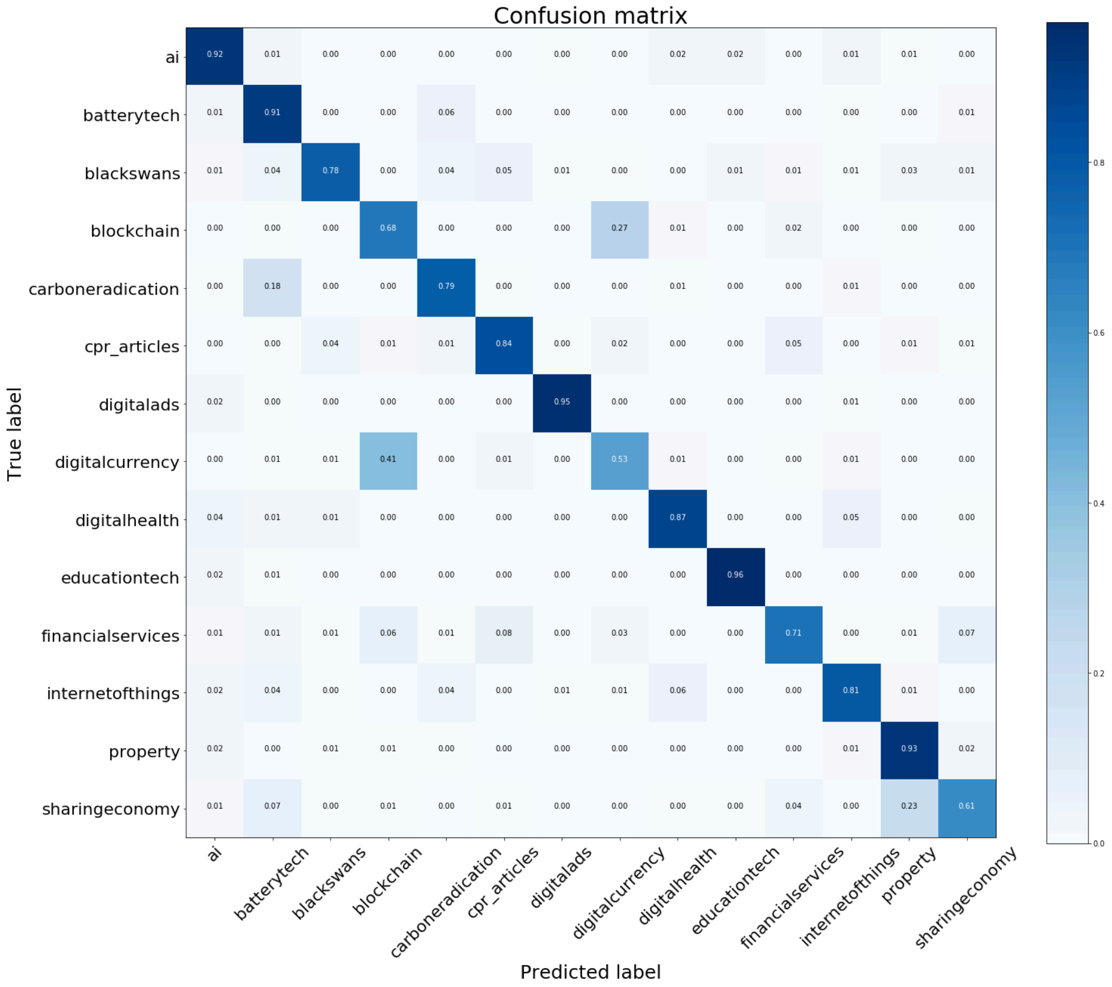
Best performance:

81.3% accuracy (balanced data)

Support vector machine benchmark:

77% accuracy

VERY DEEP CNN



	ai	batterytech	blackswans	blockchain	carboneradication	cpr_articles	digitalads	digitalcurrency	digitalhealth	educationtech	financialservices	internetofthings	property	sharingeconomy
ai	0.92	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.01	0.01	0.00
batterytech	0.01	0.91	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
blackswans	0.01	0.04	0.78	0.00	0.04	0.05	0.01	0.00	0.00	0.01	0.01	0.01	0.03	0.01
blockchain	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.27	0.01	0.00	0.02	0.00	0.00	0.00
carboneradication	0.00	0.18	0.00	0.00	0.79	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
cpr_articles	0.00	0.00	0.04	0.01	0.01	0.84	0.00	0.02	0.00	0.00	0.05	0.00	0.01	0.01
digitalads	0.02	0.00	0.00	0.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.01	0.00	0.00
digitalcurrency	0.00	0.01	0.01	0.41	0.00	0.01	0.00	0.53	0.01	0.00	0.00	0.01	0.00	0.00
digitalhealth	0.04	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.87	0.00	0.00	0.05	0.00	0.00
educationtech	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00
financialservices	0.01	0.01	0.01	0.06	0.01	0.08	0.00	0.03	0.00	0.00	0.71	0.00	0.01	0.07
internetofthings	0.02	0.04	0.00	0.00	0.04	0.00	0.01	0.01	0.06	0.00	0.00	0.81	0.01	0.00
property	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.93	0.02
sharingeconomy	0.01	0.07	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.04	0.00	0.23	0.61

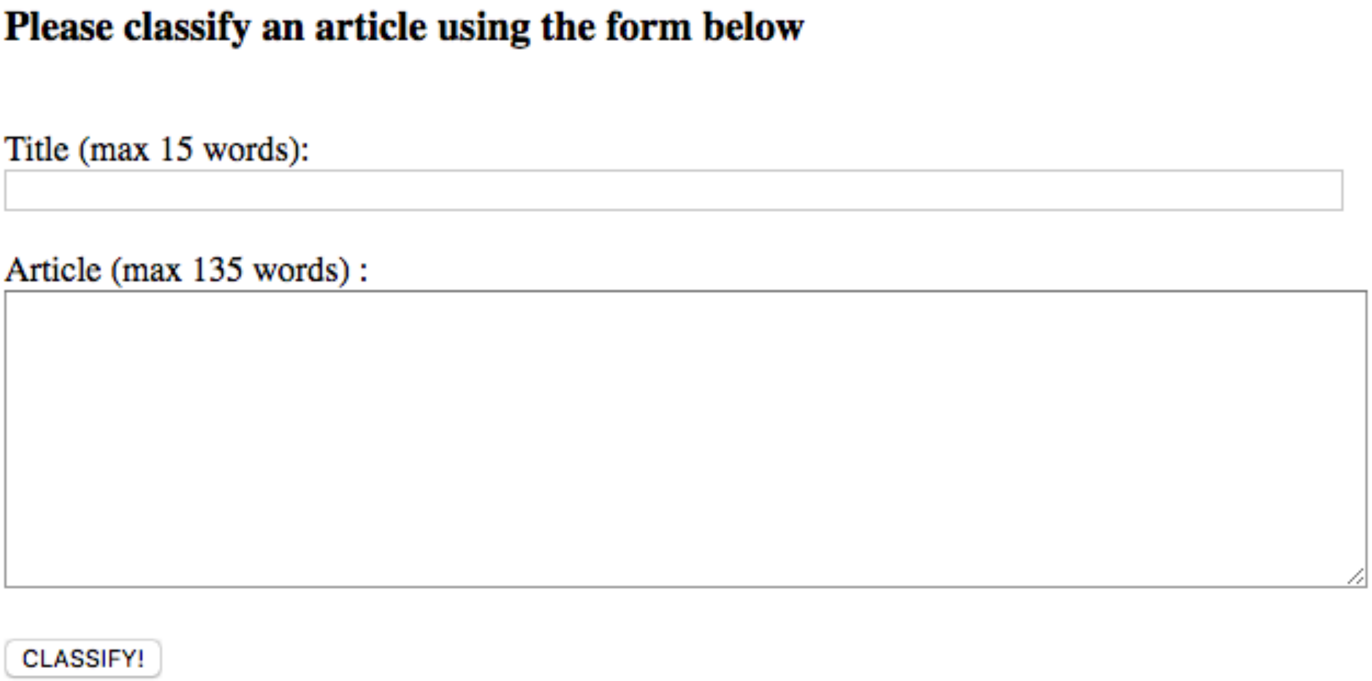
Best performance:

80.6% accuracy (balanced data)

Support vector machine benchmark:

77% accuracy

PREDICTION WEB FORM



Please classify an article using the form below

Title (max 15 words):

Article (max 135 words) :

https://afternoon-shelf-15457.herokuapp.com/form

PREDICTION API

```
import requests
import json

url = 'https://afternoon-shelf-15457.herokuapp.com/predict'
s = {"title":”foo", "body":”bar"}
s_json = json.dumps(s)
headers = {'Content-Type': 'application/json'}
r = requests.post(url, data=s_json, headers=headers)
print(r.text)
```

https://afternoon-shelf-15457.herokuapp.com/predict

NEXT STEPS

- Text Summarisation
- Tag generation
- Retrain using user feedback
- Enable batch processing

See REPORT document for details on each of these

THANKS & GOOD LUCK