# NLP on News Articles

By Edith Lee and Joleena Marshall

# Project Description

In our current attention economy, news sites are constantly competing for readers and clicks. To better understand your audience is to increase ad revenue. We've decided to do **article category classification** so news site might undertake to optimize reader retention.

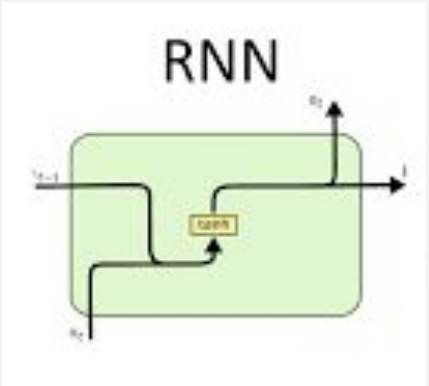
**Objective**: Classify news article category by the headline

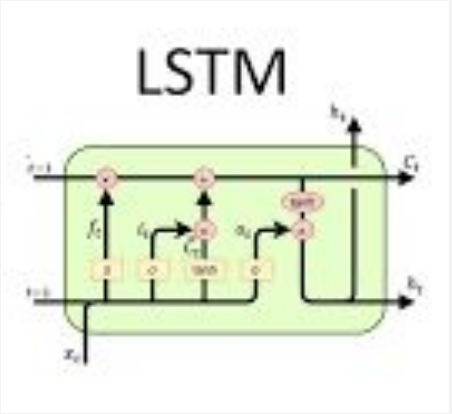
**Dataset**: CNN News Articles from 2011 to 2022 from Kaggle

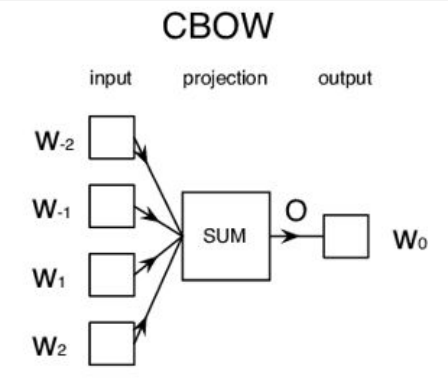
- 38,000 news articles
- Columns include: author, publication date, category, section, URL, headline, description, keywords, second headline, article text

# What We Did

- EDA to prepare for category classification
- Baseline Model: LinearSVC
- BERT (Took too long)
- CBOW Model
- RNN Model
- LSTM Model



# What We Did (bonus)

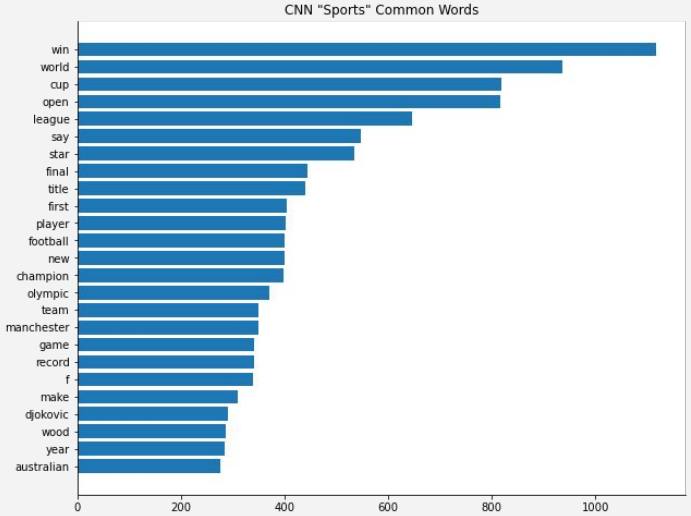
- Headline generation using LSTM

```
lstm_model.gen_seq('cup', 10)
```

```
cup
pharoah
visa
bull
youtuber
astrazeneca
pence
hacking
value
becker
flash
```

```
lstm_model.gen_seq('cup', 10)
```

```
cup
captain
terry
UNK
UNK
UNK
UNK
UNK
UNK
UNK
UNK
```



```
lstm_model.gen_seq('police', 10)
```

```
police
probe
UNK
UNK
UNK
UNK
UNK
UNK
UNK
UNK
UNK
```

```
lstm_model.gen_seq('team', 10)
```

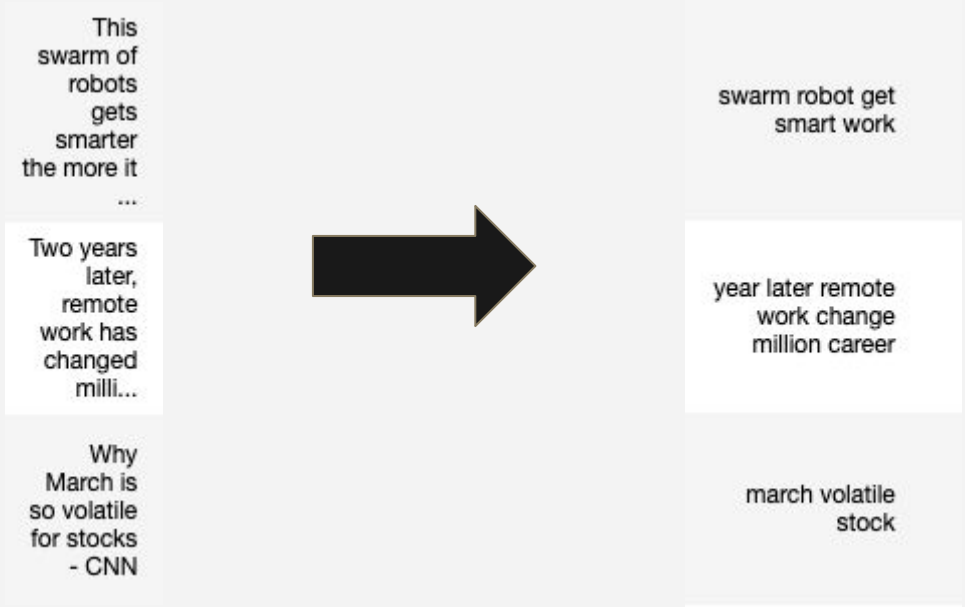
```
team
usa
UNK
UNK
UNK
UNK
UNK
UNK
UNK
UNK
UNK
```

```
lstm_model.gen_seq('london', 10)
```

```
london
police
officer
shoot
UNK
UNK
UNK
UNK
UNK
UNK
UNK
```

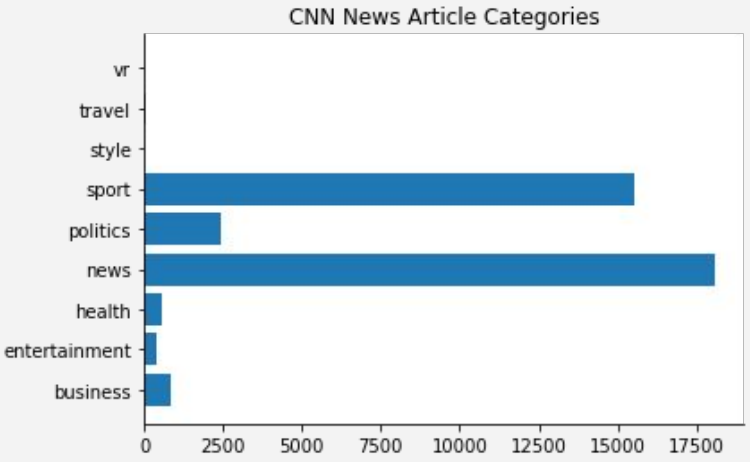
# What Went Well

- LSTM > RNN > CBOW in terms of accuracy
- Preprocessing of text made sense: “cleaned” headline looked good



# What Didn't Go Well

- BERT model took a really long time to train
- Imbalanced classes :( might have led to low model performance
- Text generation: figuring out sizes and padding, many UNK values




```
lstm_model.gen_seq('london', 10)
```

```
london
police
officer
shoot
UNK
UNK
UNK
UNK
UNK
UNK
UNK
```

# Performance Metrics

| Model | Validation Accuracy |
|---|---|
| LinearSVC | 91% |
| LSTM | 75.5% |
| RNN | 61% |
| CBOW | 47% |

# Thank you!