Fake News Detection

Tina Yuan / Pin-Tsung Huang

Overview

- Build fake news classifier using 4 different models
  - Multinomial Naive Bayes
  - Support Vector Machine
  - LSTM
  - Random Forest Classifier

- Compare performance and efficiency of different models

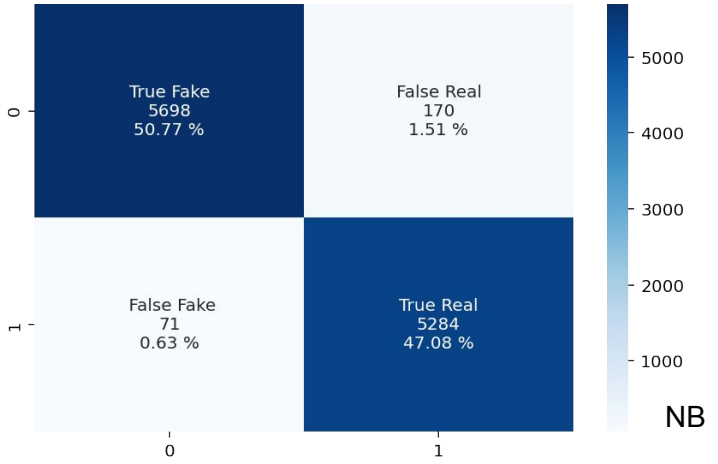
Dataset

- Articles and twitter posts that have been labeled as fake or true
- https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset
- Data has text and label columns after cleaning

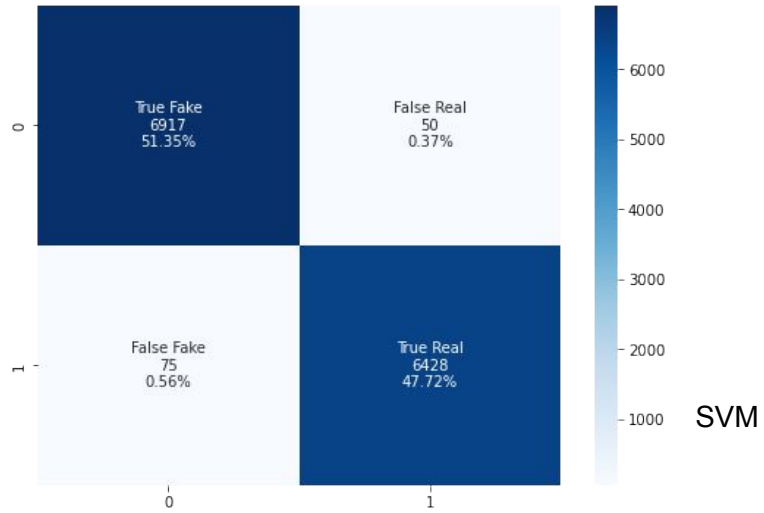
	text	label
donald trump sends embarrassing new year eve m...		0
drunk bragging trump staffer started russian c...		0
sheriff david clarke becomes internet joke thr...		0
trump obsessed even obamas name coded website ...		0
pope francis called donald trump christmas spe...		0
	...	...
fully committed nato back new u approach afgha...		1
lexisnexis withdrew two product chinese market...		1
minsk cultural hub becomes authoritiesminsk re...		1
vatican upbeat possibility pope francis visiti...		1
indonesia buy 114 billion worth russian jetsja...		1

Performance of Four Classifiers

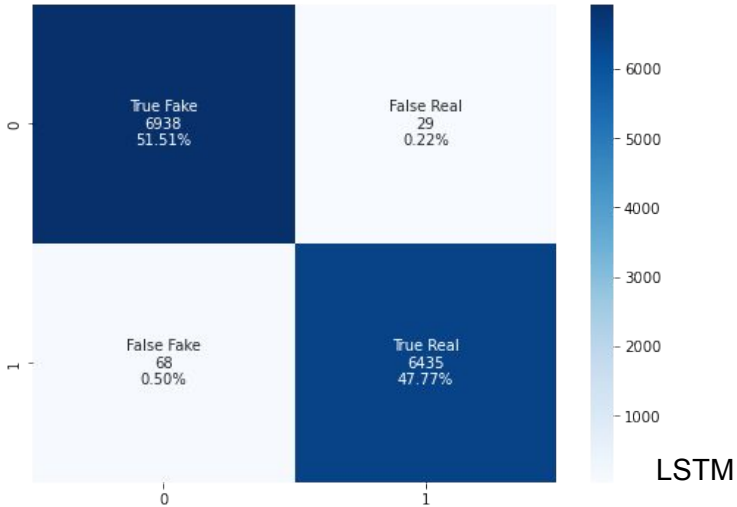
- Multinomial Naive Bayes:
  - Accuracy: 0.979
  - Training time: 0.21 sec; Inference time: 0.12 sec
- Support Vector Machine
  - Accuracy: 0.99
  - Training time: 1 min 58 sec ; Inference time: 20 sec
- LSTM Network
  - Accuracy: 0.99
  - Training time (10 epochs): 1 hr 2 min ; Inference time: 34.9 sec
- Random Forest Classifier
  - Accuracy: 0.984
  - Training time (n_estimators=120): 31.8 sec; Inference time: 0.96 sec



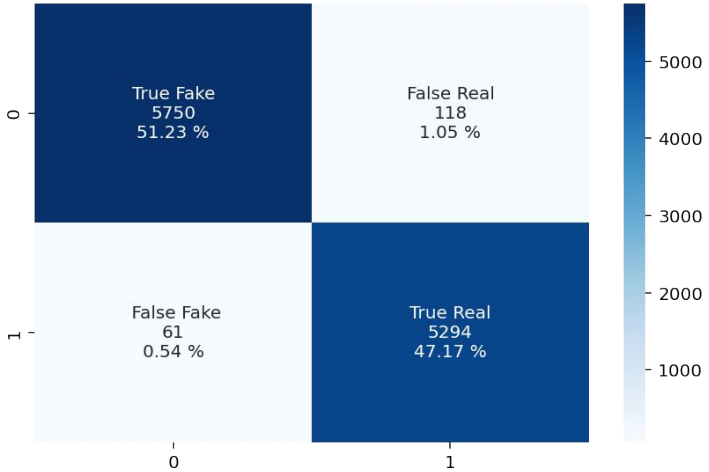
NB



SVM



LSTM



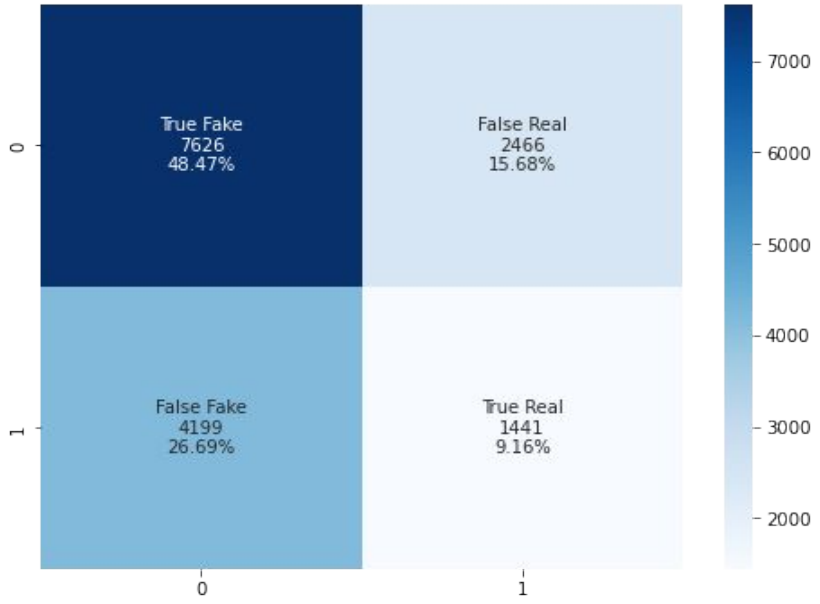
RF

On a very different dataset

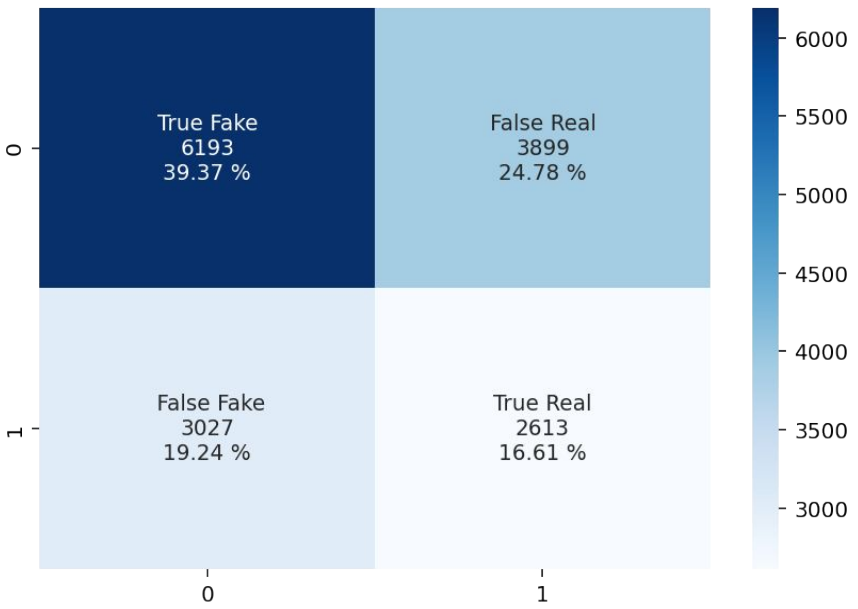
- We tested our trained models with a very different fake news dataset which contains a lot (20x) shorter paragraph
- Trained on: avg. 1776 chars
- Tested on: avg. 75 chars

Model performance on different datasets (as test data)

LSTM: Accuracy: 0.58

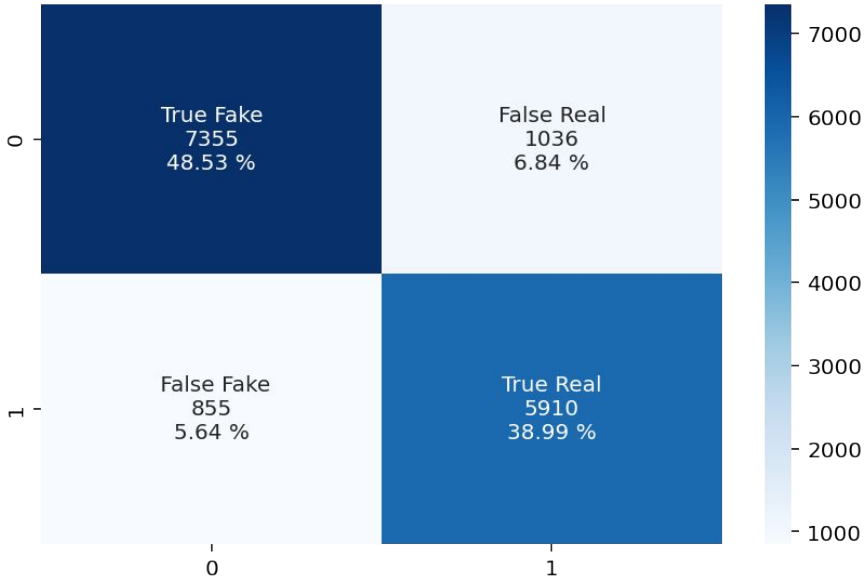


NB: Accuracy: 0.57



Combined dataset

- Training with a combined dataset (long + short) produces better results, yet the accuracy is still subpar.



NB: Accuracy: 0.88

Conclusion & Observations

- Multinomial Naive Bayes is fast and its accuracy is good enough
- LSTM is the best in terms of accuracy
- Support Vector Machine provides one of the best accuracy and acceptable training/inference time

- Data with large difference in input lengths could greatly impact the accuracy
- Applying trained models directly to other data could be almost useless