SEP 788/789 - Deep Learning and Neural Networks

Fake News Detection

Group16: Ruiqiao Wang

Zhuangyuan Shen

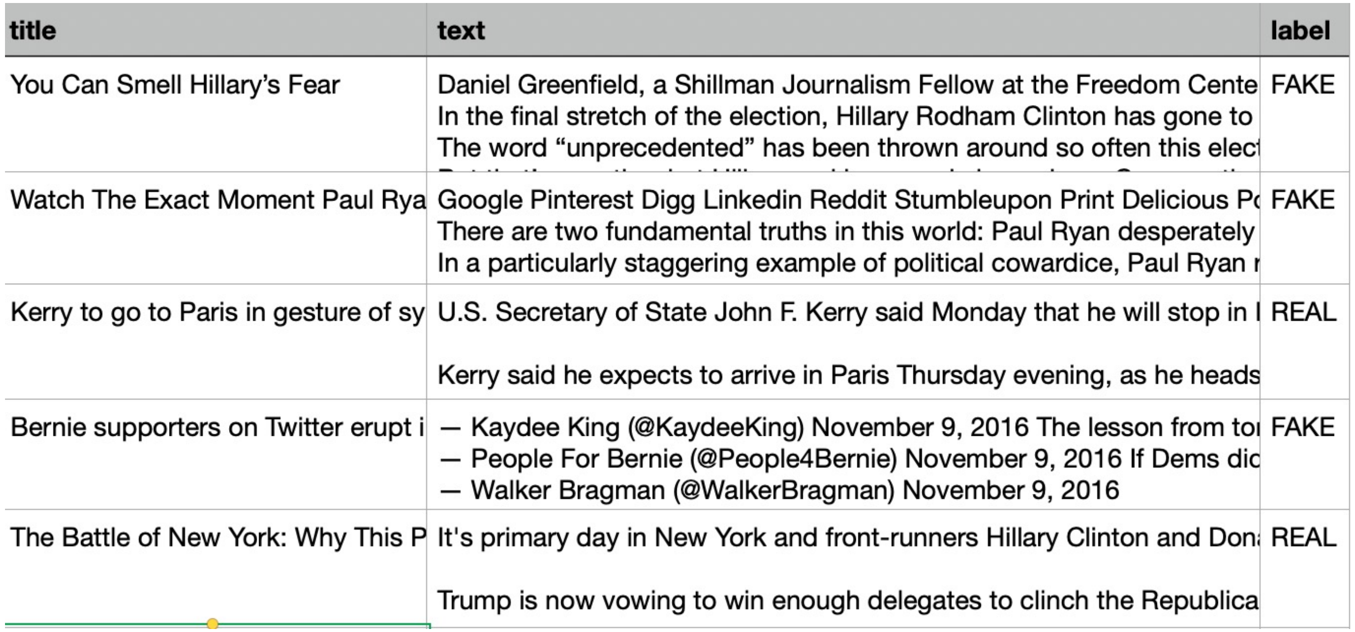
Siqi Zhao

Agenda

- Problem Statement
- Project Challenge
- Development Framework
- Proposed Approach
- Result
- Difficulties and Solutions
- Group Reflection and Improvement

Problem Statement

- Using AI to predict the likelihood of REAL news
- NLP text binary classification problem
- Dataset: 6336 pieces of news belonging to one of the two classes- REAL or FAKE



title	text	label
You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fellow at the Freedom Cente In the final stretch of the election, Hillary Rodham Clinton has gone to The word "unprecedented" has been thrown around so often this elect	FAKE
Watch The Exact Moment Paul Rya	Google Pinterest Digg Linkedin Reddit Stumbleupon Print Delicious Po There are two fundamental truths in this world: Paul Ryan desperately In a particularly staggering example of political cowardice, Paul Ryan r	FAKE
Kerry to go to Paris in gesture of sy	U.S. Secretary of State John F. Kerry said Monday that he will stop in I Kerry said he expects to arrive in Paris Thursday evening, as he heads	REAL
Bernie supporters on Twitter erupt i	— Kaydee King (@KaydeeKing) November 9, 2016 The lesson from to — People For Bernie (@People4Bernie) November 9, 2016 If Dems dic — Walker Bragman (@WalkerBragman) November 9, 2016	FAKE
The Battle of New York: Why This P	It's primary day in New York and front-runners Hillary Clinton and Don Trump is now vowing to win enough delegates to clinch the Republica	REAL

Project Challenge

- Huge dimension of the input features - Curse of dimensionality
- Small amount of data – Only 6336 pieces of data points

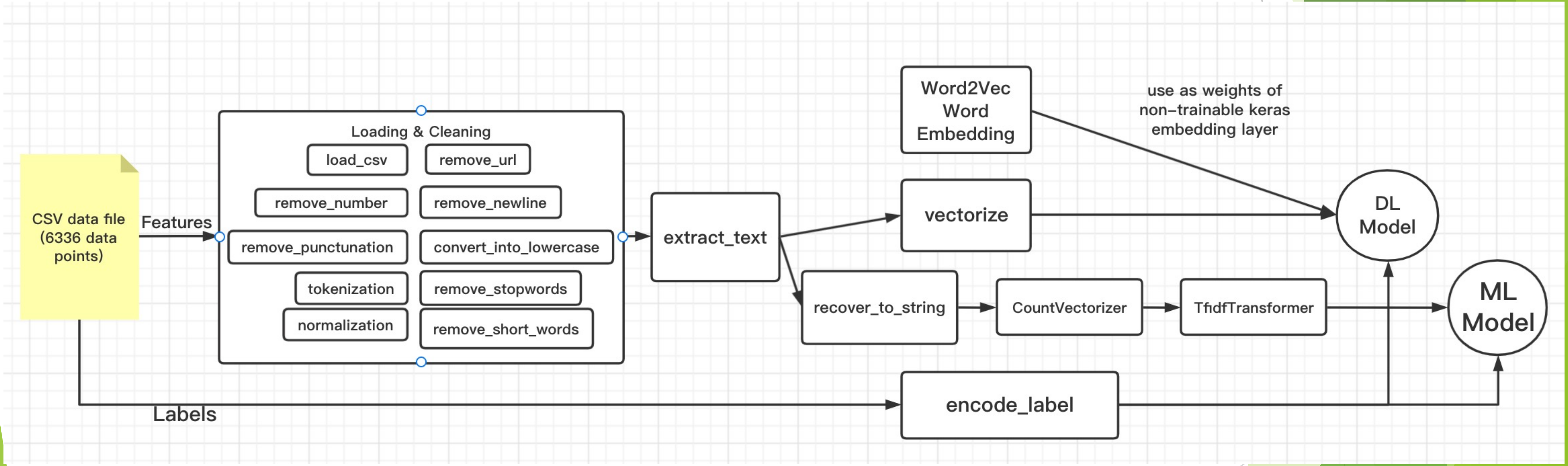
Development Framework

- NLTK
- Scikit-learn
- TensorFlow, Keras
- gensim
- Other Scientific Computing Library – NumPy, Matplotlib…

Proposed Approach

- Outline
- Data Analysis
- Preprocessing of dataset
- Machine Learning Model
- Deep Learning Model
- Metrics

Proposed Approach Outline



Data Analysis

Dataset Visualization

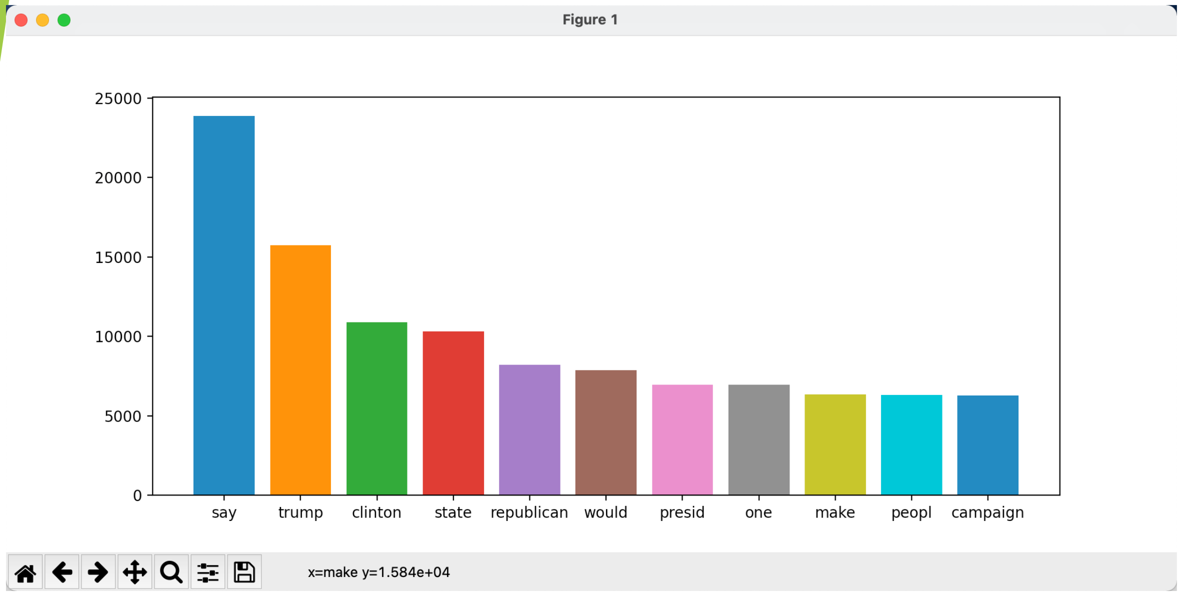


Fig1. Word frequency distribution histogram for REAL news

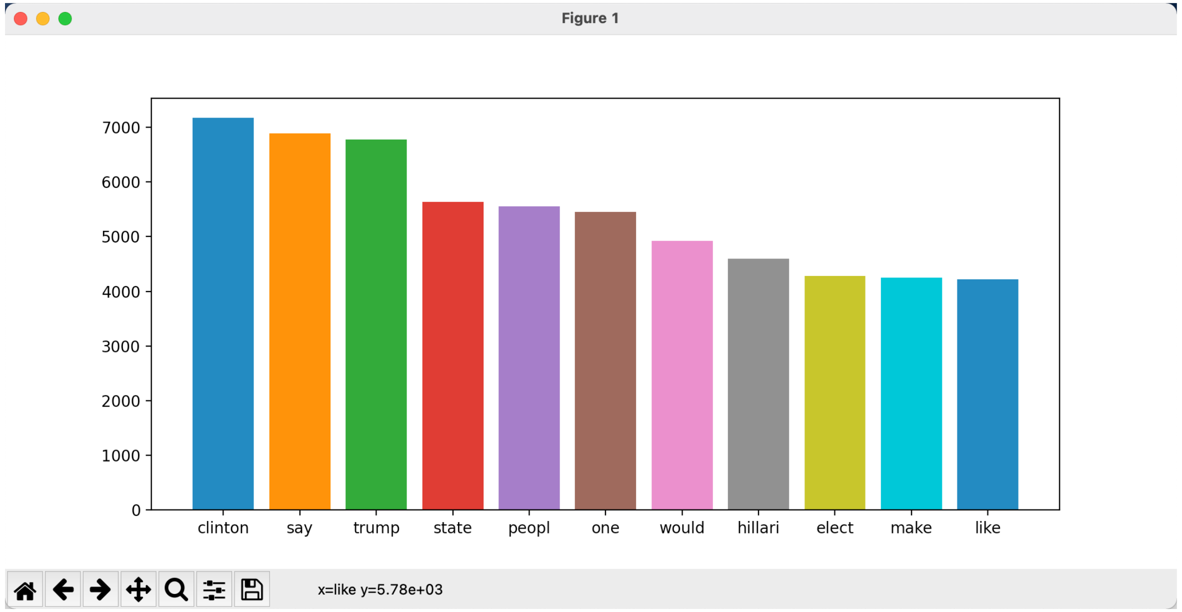
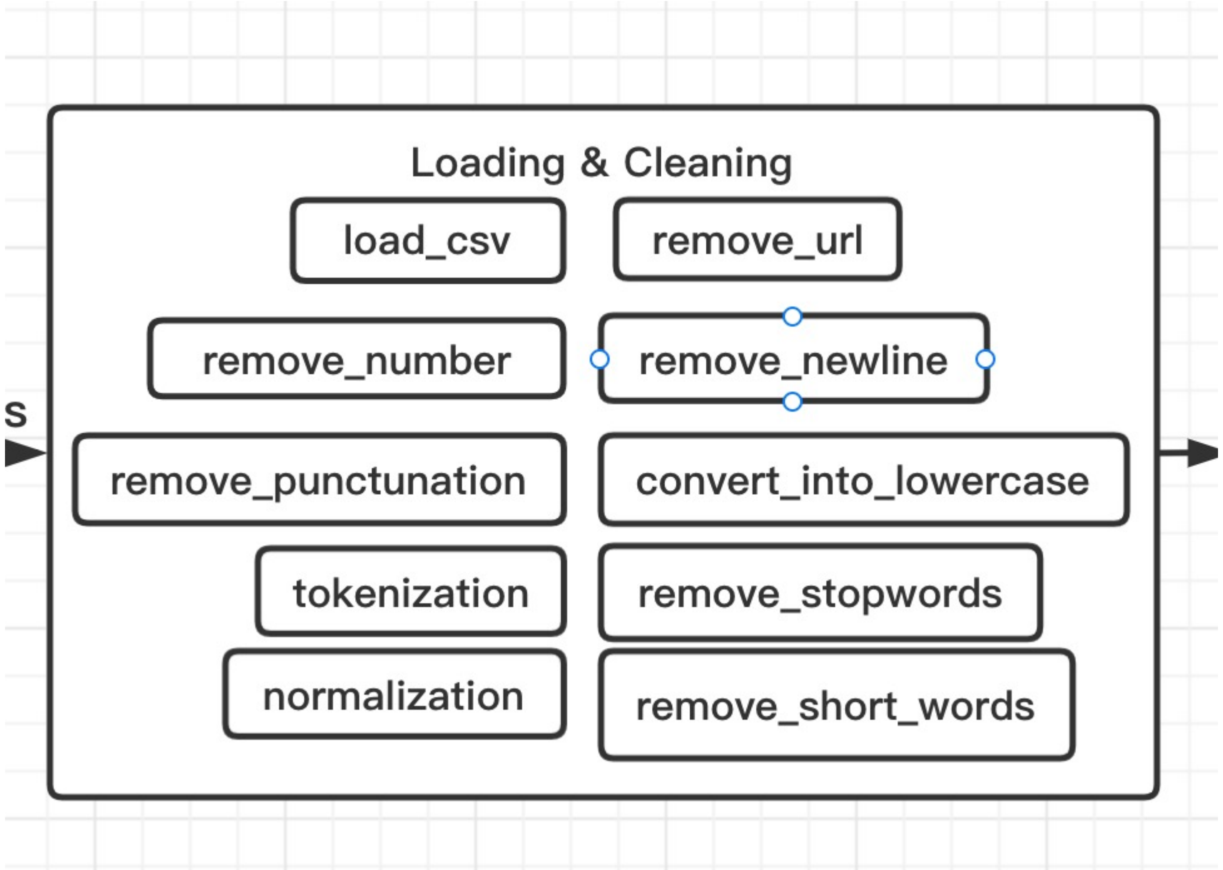


Fig2. Word frequency distribution histogram for FAKE news

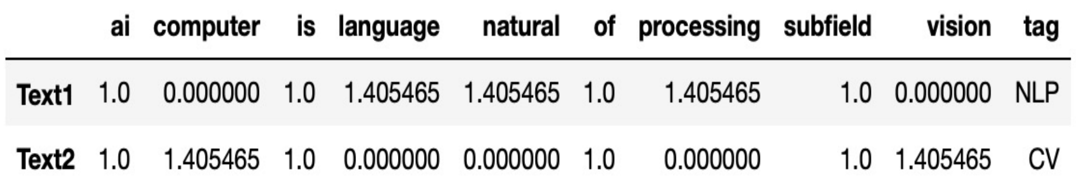
Preprocessing of dataset



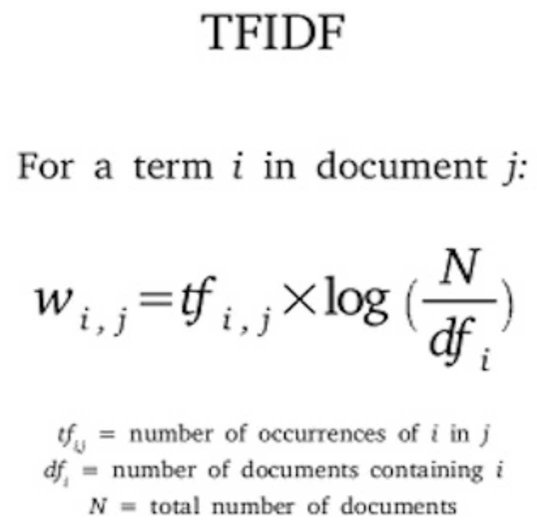
- We used regular expressions, and the NLTK language processing library to clean the data
- We can add or ignore some of these methods according to the different needs of the training model stage

Text1 = “Natural Language Processing is a subfield of AI”

Text2 = “Computer Vision is a subfield of AI”



	ai	computer	is	language	natural	of	processing	subfield	vision	tag
Text1	1.0	0.000000	1.0	1.405465	1.405465	1.0	1.405465	1.0	0.000000	NLP
Text2	1.0	1.405465	1.0	0.000000	0.000000	1.0	0.000000	1.0	1.405465	CV



TFIDF

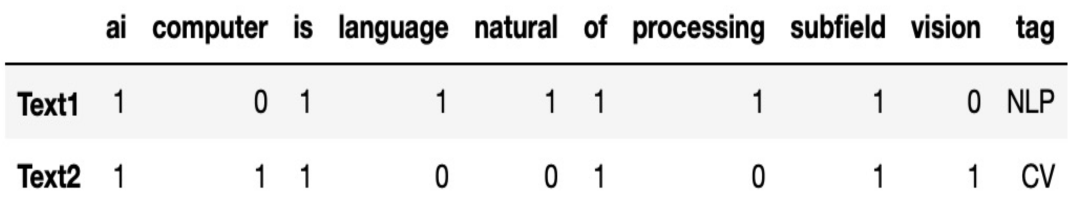
For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j

df_i = number of documents containing i

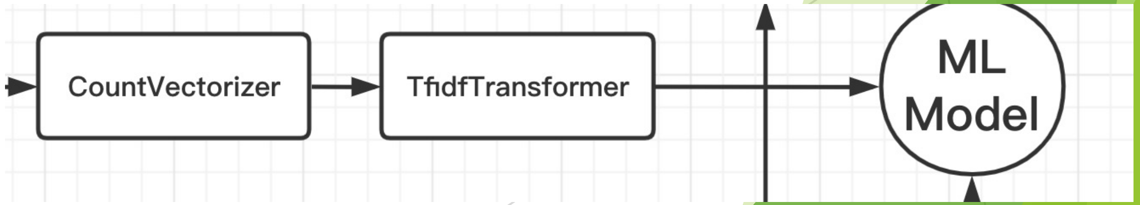
N = total number of documents



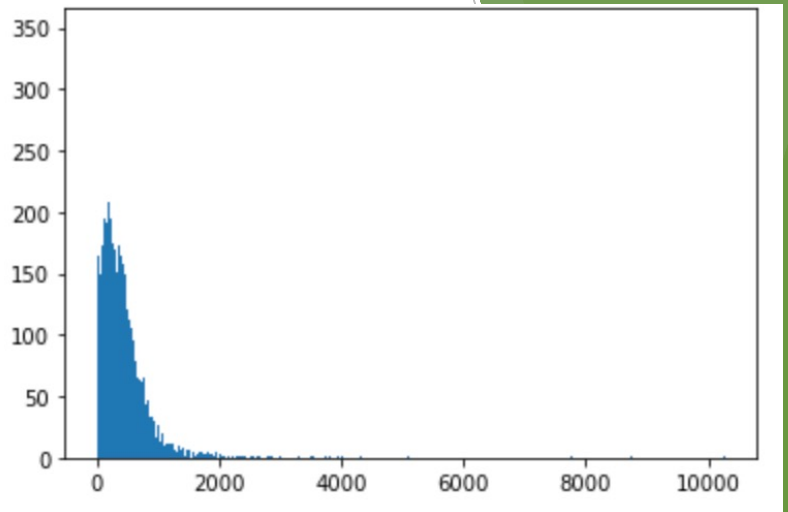
	ai	computer	is	language	natural	of	processing	subfield	vision	tag
Text1	1	0	1	1	1	1	1	1	0	NLP
Text2	1	1	1	0	0	1	0	1	1	CV

Machine Learning Model

- Build Scikit-learn Pipeline
- CountVectorizer()
- TfidfTransformer()
- Training model:
  - Naive Bayes
  - K-Nearest Neighbours
  - Support Vector Machines
  - Logistic Regression
  - Decision Tree



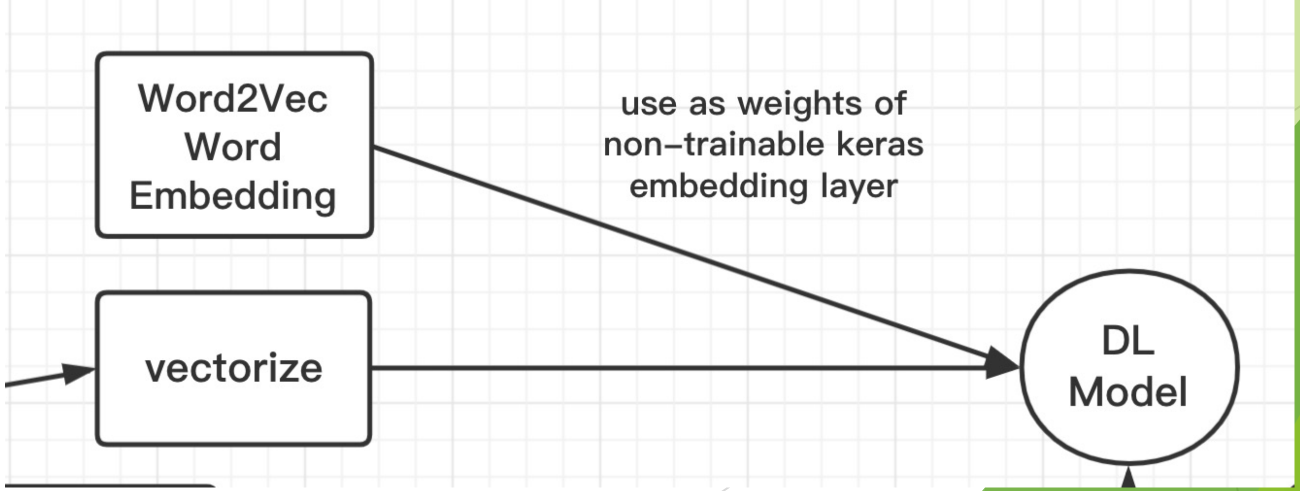
Deep Learning Model



- Implement by Keras
- Using keras' built-in Tokenizer to vectorize the text
- Padding the text
- Word embedding: Word2Vec implement by Word2Vec model of genism

Getting embedding vectors from word2vec and using its as weights of non-trainable keras embedding layer

- Early Stop to prevent over fitting
- Training model:
  - MLP
  - CNN
  - LSTM

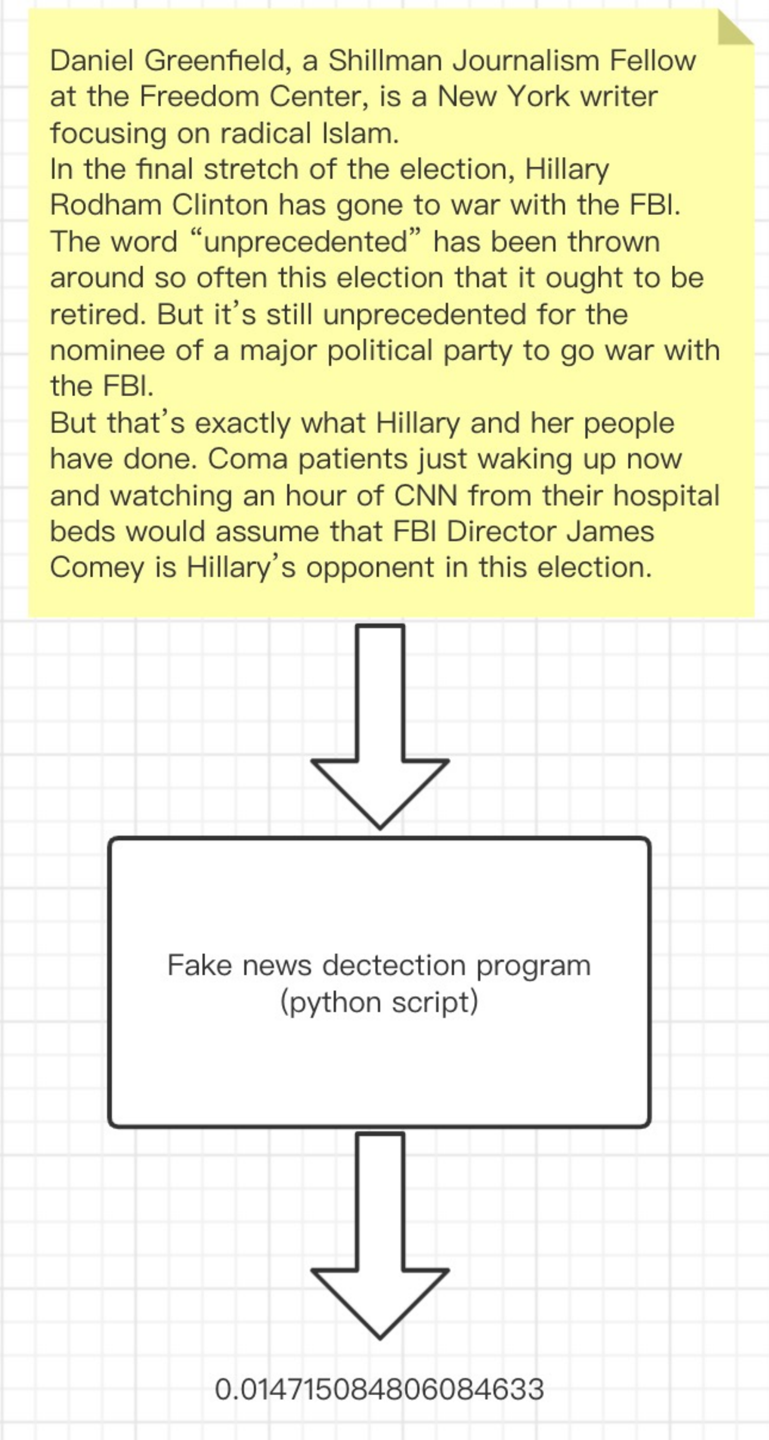


Metrics

- Precision
- Recall
- F1-Sorce
- Accuracy
- Confusion matrix

Result

Model	Accuracy
Naive Bayes	0.8333333333333334
K-Nearest Neighbours	0.8698412698412699
Support Vector Machines	0.9333333333333333
Logistic Regression	0.9317460317460318
Decision Tree	0.8047619047619048
MLP	0.9009523582458496
CNN	0.8876190185546875
LSTM	0.9141269612312317



Daniel Greenfield, a Shillman Journalism Fellow at the Freedom Center, is a New York writer focusing on radical Islam.

In the final stretch of the election, Hillary Rodham Clinton has gone to war with the FBI.

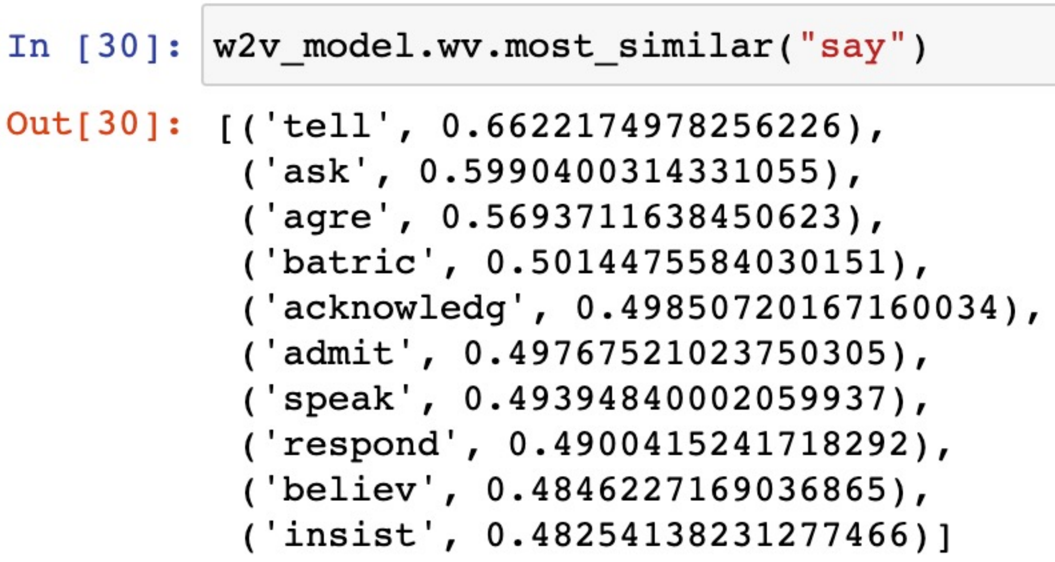
The word “unprecedented” has been thrown around so often this election that it ought to be retired. But it’s still unprecedented for the nominee of a major political party to go war with the FBI.

But that’s exactly what Hillary and her people have done. Coma patients just waking up now and watching an hour of CNN from their hospital beds would assume that FBI Director James Comey is Hillary’s opponent in this election.

0.014715084806084633

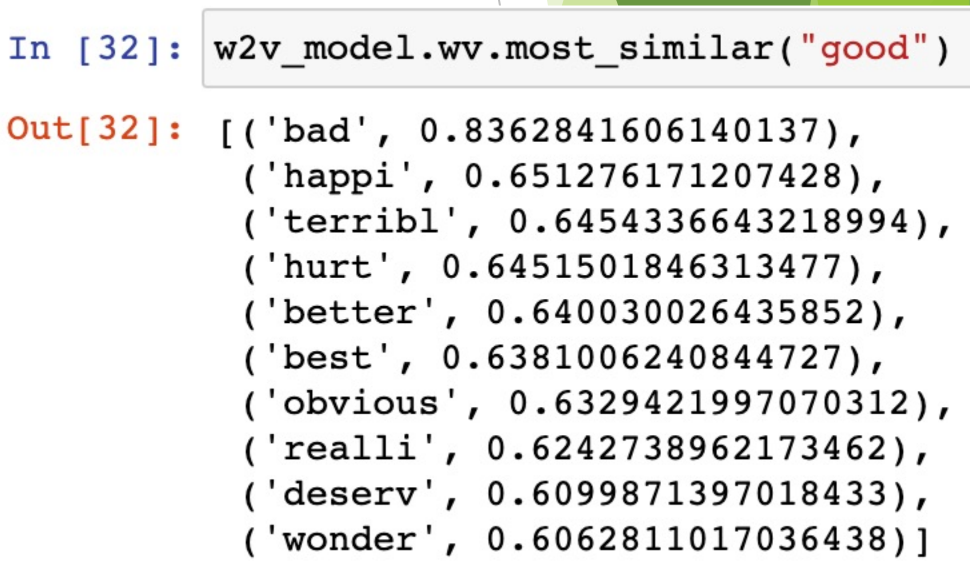
Difficulties and Solutions

- Huge dimension of the input features, lead to overfitting
  - Word2Vec word embedding, Keras Embedding layer
  - Scikit-learn CountVectorizer() & TfidfTransformer()
  - capture the similarities between two words
- Lack of data
  - Cross-Validation



```
In [30]: w2v_model.wv.most_similar("say")

Out[30]: [('tell', 0.6622174978256226),
          ('ask', 0.5990400314331055),
          ('agre', 0.5693711638450623),
          ('batric', 0.5014475584030151),
          ('acknowledg', 0.4985072016716034),
          ('admit', 0.49767521023750305),
          ('speak', 0.49394840002059937),
          ('respond', 0.4900415241718292),
          ('believ', 0.4846227169036865),
          ('insist', 0.48254138231277466)]
```



```
In [32]: w2v_model.wv.most_similar("good")

Out[32]: [('bad', 0.8362841606140137),
          ('happi', 0.651276171207428),
          ('terribl', 0.6454336643218994),
          ('hurt', 0.6451501846313477),
          ('better', 0.640030026435852),
          ('best', 0.6381006240844727),
          ('obvious', 0.6329421997070312),
          ('realli', 0.6242738962173462),
          ('deserv', 0.6099871397018433),
          ('wonder', 0.6062811017036438)]
```

Improvement or Future Work

- NLP Data Augmentation

  - Increase the amount of training data and improve the generalization ability of the model

  - Add noise data to improve the robustness of the model

- Try more modern model: Transformer/Google Bert

  - Solve the lack of data problem

Group Reflection

- Practice and made us more familiar with development using machine learning or deep learning frameworks
- Practice the NLP problem, improve experience

Thank you for watching

Group16: Ruiqiao Wang
Zhuangyuan Shen
Siqi Zhao