Fake News Detection with Text CNN

Yiyang Sun, Handong Xu

Mentor: Prof. Ratan Dey
Superviosr: Prof. Olivier Gilles Marin

NYU Shanghai Computer Science Capstone

May 2019

Outline

The Popularization of Fake News

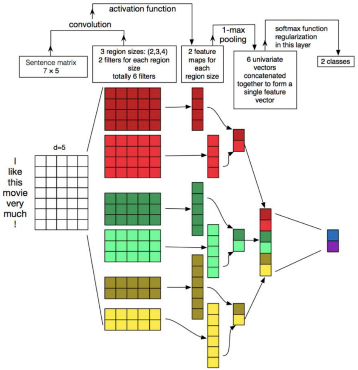
- Since the 2016 U.S. presidential election, people are more aware of the fake news on social media.
- In general fake news receive more interaction on social media like Facebook or Tweeter.

Early Stage Detection

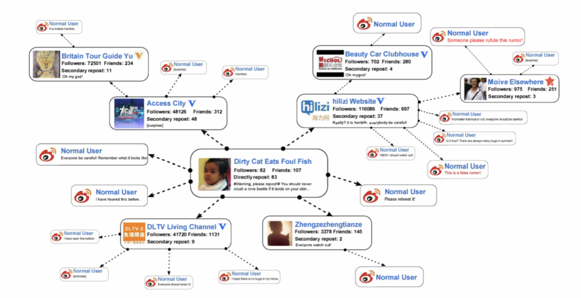
- predict based on the content of the news

Early Stage Detection

- predict based on the content of the news
- find the latent features of fake news

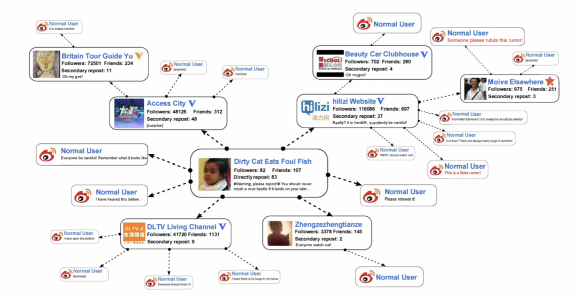


Propagation Detection



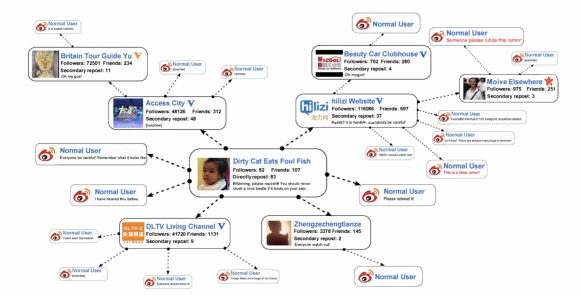
- use the information from the post's distribution route online over time to classify fake news

Propagation Detection



- use the information from the post's distribution route online over time to classify fake news
- fake news has a different propagation route than real news

Propagation Detection



- use the information from the post's distribution route online over time to classify fake news
- fake news has a different propagation route than real news
- usually assign a credit evaluation to certain users

Dataset Description

	text	title	type
0	They stood in line at Trump Tower, sometimes u...	At Donald Trump<U+2019>s Properties, a Showcas...	1
1	Donald J. Trump <U+2019>s foundation informed ...	Trump Foundation Tells New York It Has Stopped...	1
2	President-elect Donald J. Trump won the White ...	Donald Trump Prepares for White House Move, bu...	1
3	An investment pitch for a new Texas hotel is t...	Luring Chinese Investors With Trump<U+2019>s N...	1
4	President-elect Donald J. Trump <U+2019>s wife...	Melania and Barron Trump Won<U+2019>t Immediat...	1

- Data Attributes: title, text, label(0-fake,1-real)

Dataset Description

	text	title	type
0	They stood in line at Trump Tower, sometimes u...	At Donald Trump<U+2019>s Properties, a Showcas...	1
1	Donald J. Trump <U+2019>s foundation informed ...	Trump Foundation Tells New York It Has Stopped...	1
2	President-elect Donald J. Trump won the White ...	Donald Trump Prepares for White House Move, bu...	1
3	An investment pitch for a new Texas hotel is t...	Luring Chinese Investors With Trump<U+2019>s N...	1
4	President-elect Donald J. Trump <U+2019>s wife...	Melania and Barron Trump Won<U+2019>t Immediat...	1

- Data Attributes: title, text, label(0-fake,1-real)
- 40% fake news from kaggle + 60% real news from trusted website like The New York Times.

Dataset Description

	text	title	type
0	They stood in line at Trump Tower, sometimes u...	At Donald Trump<U+2019>s Properties, a Showcas...	1
1	Donald J. Trump <U+2019>s foundation informed ...	Trump Foundation Tells New York It Has Stopped...	1
2	President-elect Donald J. Trump won the White ...	Donald Trump Prepares for White House Move, bu...	1
3	An investment pitch for a new Texas hotel is t...	Luring Chinese Investors With Trump<U+2019>s N...	1
4	President-elect Donald J. Trump <U+2019>s wife...	Melania and Barron Trump Won<U+2019>t Immediat...	1

- Data Attributes: title, text, label(0-fake,1-real)
- 40% fake news from kaggle + 60% real news from trusted website like The New York Times.
- It only covers news on US presidential election from October 2016 to November 2016.

Data Cleaning

- select only the texts written in English since different languages contain different structures of writing

Data Cleaning

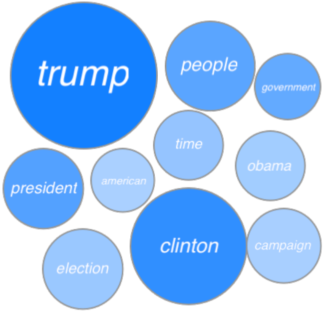
- select only the texts written in English since different languages contain different structures of writing
- filter out the commonly used stop words like ‘is’, ‘it’

Data Cleaning

- select only the texts written in English since different languages contain different structures of writing
- filter out the commonly used stop words like ‘is’, ‘it’
- filter out special symbols like ‘#’, ‘@’

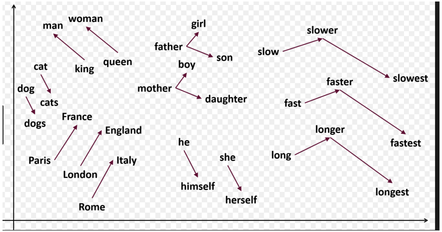
Data Processing - Word Frequency Dictionary

- count the word frequency in the dataset and sort in descending order
- map the word to its position in the dictionary
- easy and efficient to implement, we can add embedding layers to the vectors afterwards

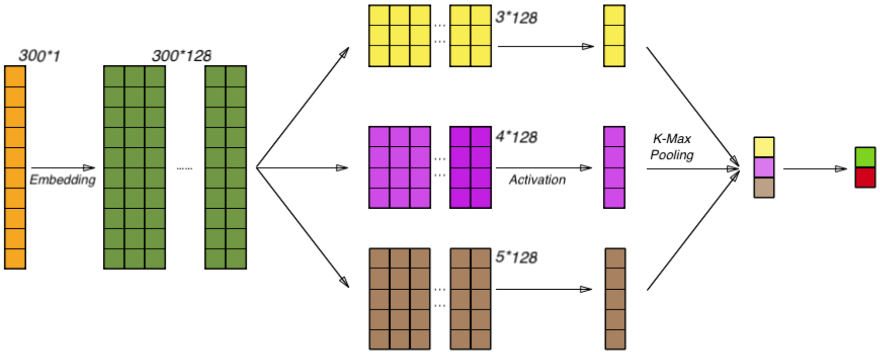


Data Processing - Word2vec

- a deep learning method to turn word into vectors
- map words with similar meanings together
- takes more time to run and sensitive to parameters

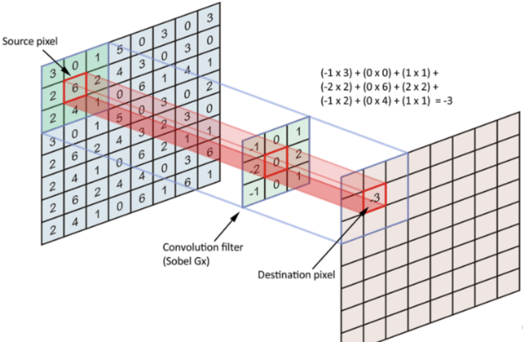


Model Architecture-textCNN



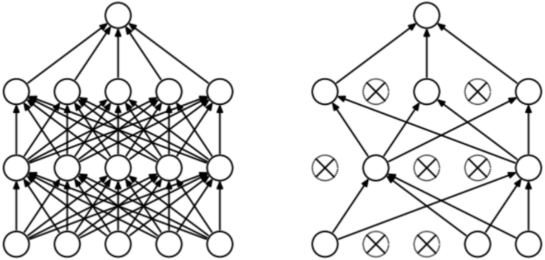
Hyperparamters and Training

- use the frequency-based vectors as input
- add embedding layers to amplify the information
  $\rightarrow$ *number of records*, *sequence length*, 128
- filter size of 3*128, 4*128,5*128, we choose size x*128 so that the filter only moves up and down



Hyperparamters and Training

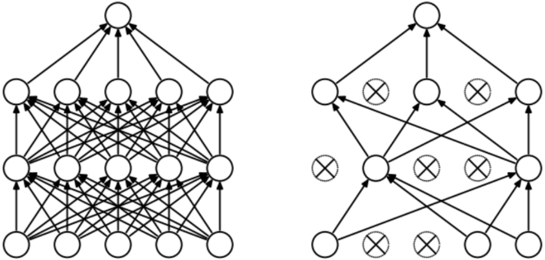
- dropout rate of 0.5 to avoid over fitting



(a) Standard Neural Net (b) After applying dropout.

Hyperparamters and Training

- dropout rate of 0.5 to avoid over fitting

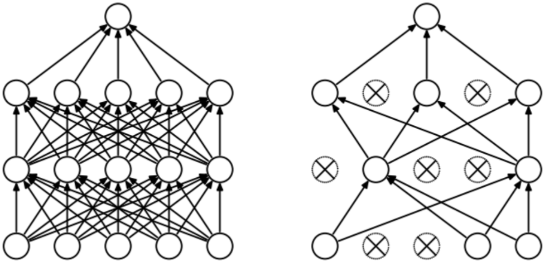


(a) Standard Neural Net (b) After applying dropout.

- batch size of 200, so each time 200 training samples get trained

Hyperparamters and Training

- dropout rate of 0.5 to avoid over fitting



(a) Standard Neural Net (b) After applying dropout.

- batch size of 200, so each time 200 training samples get trained
- use binary entropy loss since its a binary classification problem (real or fake) and Adam optimizer

Other Machine Learning Algorithms

- Logistic Regression with frequency based vectors and w2v vectors

Other Machine Learning Algorithms

- Logistic Regression with frequency based vectors and w2v vectors
- Support Vector Machine with Gaussian kernel with frequency based vectors and w2v vectors

Other Machine Learning Algorithms

- Logistic Regression with frequency based vectors and w2v vectors
- Support Vector Machine with Gaussian kernel with frequency based vectors and w2v vectors
- Random Forest with frequency based vectors and w2v vectors

Other Machine Learning Algorithms

- Logistic Regression with frequency based vectors and w2v vectors
- Support Vector Machine with Gaussian kernel with frequency based vectors and w2v vectors
- Random Forest with frequency based vectors and w2v vectors
- Boosted Tree with frequency based vectors and w2v vectors

Other Machine Learning Algorithms

- Logistic Regression with frequency based vectors and w2v vectors
- Support Vector Machine with Gaussian kernel with frequency based vectors and w2v vectors
- Random Forest with frequency based vectors and w2v vectors
- Boosted Tree with frequency based vectors and w2v vectors
- Long Short Term Memory with frequency based vectors

Ensemble Different Algorithms

- To get better result, we try to ensemble the results from two different algorithms

Ensemble Different Algorithms

- To get better result, we try to ensemble the results from two different algorithms
- We choose SVM and logistic regression because SVM returns 0/1, while LR return the probability of the records being true

Ensemble Different Algorithms

- To get better result, we try to ensemble the results from two different algorithms
- We choose SVM and logistic regression because SVM returns 0/1, while LR return the probability of the records being true
- textCNN + freq_SVM
- textCNN + w2v_SVM
- textCNN + freq_Logistic_regression
- textCNN + w2v_Logistic_regression

AUC Value and Run Time

Accuracy	Dataset 1	Running Time(s)	Dataset 2	Running Time(s)
freq-LR	64.99%	0.97	61.81%	0.04
freq-SVM	50.25%	13.66	53.66%	0.06
freq-RandomForest	70.34%	3.91	61.96%	0.26
freq-BoostedTree	95.11%	25.31	68.19%	0.24
freq-RNN	99.11%	4065	62.08%	139
freq-textCNN	98.84%	448	50.00%	13
w2v-LR	88.98%	0.52	78.54%	0.08
w2v-SVM	89.15%	222.85	78.69%	0.11
w2v-RandomForest	83.09%	2.97	75.79%	0.19
w2v-BoostedTree	89.45%	28.87	71.06%	0.49
w2v-RNN	Not Applicable			
w2v-textCNN	Not Applicable			
*w2v = word2vector *freq = frequency dictionary				

Precision, Recall and F1 stats

Accuracy	Dataset 1			Dataset 2		
	Precision	Recall	F1	Precision	Recall	F1
freq-LR	0.600	0.421	0.495	0.667	0.553	0.605
freq-SVM	1	0.005	0.010	0.553	1.0	0.712
freq-RandomForest	0.790	0.460	0.581	0.676	0.532	0.595
freq-BoostedTree	0.975	0.912	0.942	0.711	0.681	0.696
freq-RNN	0.99	0.99	0.99	0.857	0.222	0.353
freq-textCNN	0.994	0.987	0.990	0.628	1.000	0.771
w2v-LR	0.862	0.838	0.850	0.818	0.766	0.791
w2v-SVM	0.860	0.842	0.851	0.833	0.745	0.787
w2v-RandomForest	0.854	0.715	0.778	0.760	0.809	0.784
w2v-BoostedTree	0.863	0.847	0.855	0.712	0.787	0.747
w2v-RNN	Not Applicable					
w2v-textCNN	Not Applicable					

*w2v = word2vector *freq = frequency dictionary

First Phase Analysis

- textCNN gives a balanced prediction result, while others like SVM predict most of the test set to be mostly real of mostly fake

First Phase Analysis

- textCNN gives a balanced prediction result, while others like SVM predict most of the test set to be mostly real of mostly fake
- textCNN and LSTM are similar in test accuracy, but textCNN is 10 times faster.

First Phase Analysis

- textCNN gives a balanced prediction result, while others like SVM predict most of the test set to be mostly real of mostly fake
- textCNN and LSTM are similar in test accuracy, but textCNN is 10 times faster.
- for traditional machine learning algorithms, w2v method gives a better result.

AUC Value Table

Accuracy	Without ensemble - DataSet1	With ensemble - DataSet1	Without ensemble - DataSet2	With ensemble - DataSet2
*0.9*freq-textCNN+0.1*freq-SVM*	98.84%/50.25%	98.63%	50.00%/53.66%	51.21%
*0.9*freq-textCNN+0.1*w2v-SVM*	98.84%/89.15%	98.68%	50.00%/78.69%	50.00%
*0.6*freq-textCNN+0.4*w2v-LR*	98.84%/88.98%	98.58%	50.00%/78.54%	77.16%
*0.6*freq-textCNN+0.4*freq-LR*	98.84%/64.99%	98.63%	50.00%/61.81%	61.80%

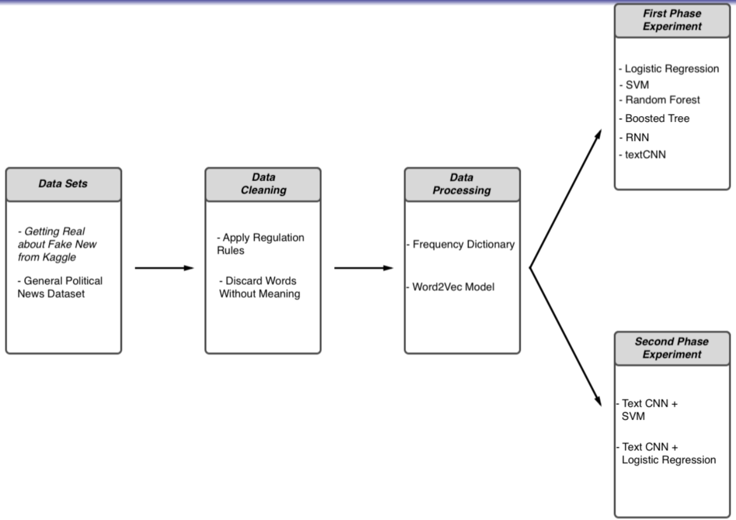
Second Phase Analysis

- ensemble gives a worse result than running one of the algorithms

Second Phase Analysis

- ensemble gives a worse result than running one of the algorithms
- possible explanations:
  - SVM returns only 0 and 1,but textCNN returns the probability of the record being true.
  - Logistic Regression is the basic version of Neural Network, so its won't be as accurate as textCNN.

Conclusion


Data Sets
- Getting Real about Fake New from Kaggle
- General Political News Dataset
Data Cleaning
- Apply Regulation Rules
- Discard Words Without Meaning
Data Processing
- Frequency Dictionary
- Word2Vec Model
First Phase Experiment
- Logistic Regression
- SVM
- Random Forest
- Boosted Tree
- RNN
- textCNN
Second Phase Experiment
- Text CNN + SVM
- Text CNN + Logistic Regression

Conclusion

- text CNN has a higher accuracy and auc value than traditional machine learning algorithms no matter the input comes from word2vec or frequency dictionary.

Conclusion

- text CNN has a higher accuracy and auc value than traditional machine learning algorithms no matter the input comes from word2vec or frequency dictionary.
- textCNN is roughly 10 times faster than LSTM, but they have a similar precision and recall statistics

Conclusion

- text CNN has a higher accuracy and auc value than traditional machine learning algorithms no matter the input comes from word2vec or frequency dictionary.
- textCNN is roughly 10 times faster than LSTM, but they have a similar precision and recall statistics
- ensemble does not help improve the result regardless of the dataset size

Conclusion

- text CNN has a higher accuracy and auc value than traditional machine learning algorithms no matter the input comes from word2vec or frequency dictionary.
- textCNN is roughly 10 times faster than LSTM, but they have a similar precision and recall statistics
- ensemble does not help improve the result regardless of the dataset size
- word2vec prepossessing leads to a better result on small datasets

Future Works

- Crawl more data on the new US president election

Future Works

- Crawl more data on the new US president election
- Get propagation structure and data

Future Works

- Crawl more data on the new US president election
- Get propagation structure and data
- Enable propagation stage detection after we finish early stage text detection

References

Y. Kim. "Convolutional Neural Networks for Sentence Classification" *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.* (October, 2014): 1746-1751

Yang, Zheng, Zhang, Cui, Li and Philip S. Yu. "TI-CNN: Convolutional Neural Networks for Fake News Detection. " *CoRR* abs/1806.00749 (August 13, 2018)

Thank You

Thank You!