Dimitris Glikiotis

Weak Supervision for Fake News Detection via Reinforcement Learning

Abstract

- Social media is one of the main ways of information. More than 68% of American adults get news on social media at least occasionally.
- Fake news: Intentionally and verifiably false news stories
- The spread of fake news may bring many negative impacts, including social panic and financial loss.
- There is a need for identifying and stopping the spreading of the fake news in social media platforms.

Most common ways to identify fake news

- Traditional learning models.
  - Typically extract features from news articles and train classifiers based on the extracted features.
- Deep learning models.
  - + improvement in the performance of fake news detection due to their powerful abilities of learning informative representations.
  - - Requires a large amount of hand-labeled data (Expensive)
  - - Data can be quickly outdated and irrelevant
  - - Lack of fresh high-quality samples for training

  Possible solution for input data**: Leverage the feedback provided by users who read the news.**

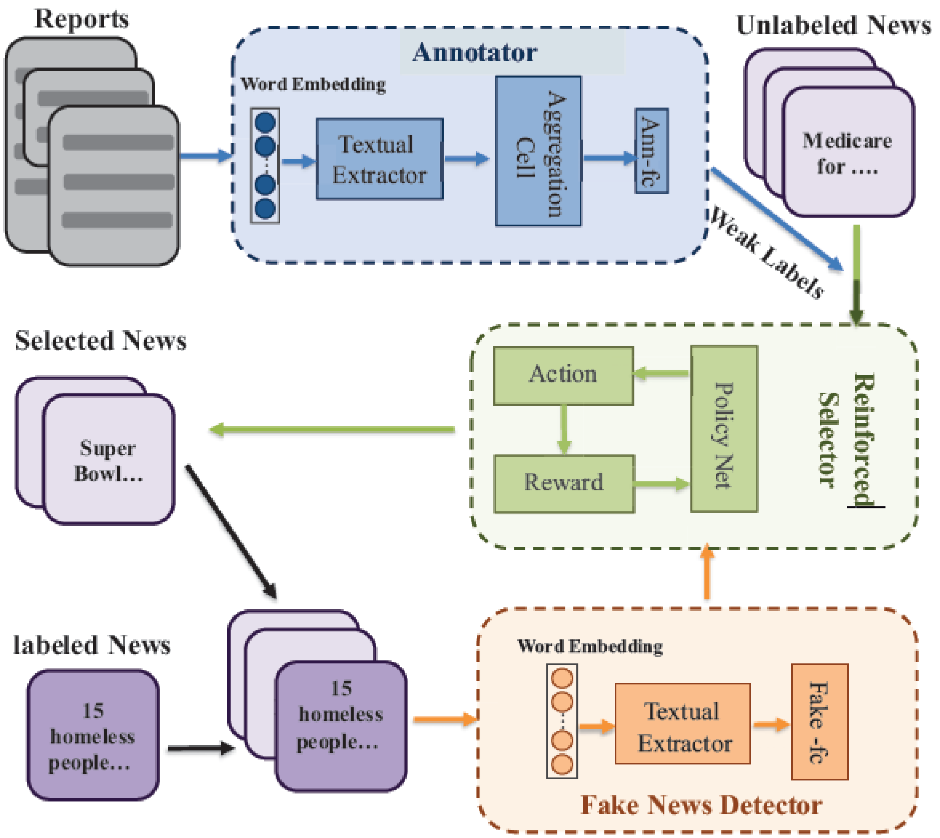
Proposal: WEakly-supervised FakE News Detection framework

(WeFEND)

- Problem: Label shortage issue -> Proposal: Leverage user reports as weak supervision
- WeFEND framework can automatically annotate articles, which help enlarge the size of the training set and this means success of deep learning models in fake news detection
- Reinforcement learning techniques: WeFEND has the ability of selecting high-quality samples.
- The article empirically shows that WeFEND can identify fake news and outperform other fake news detection models

Methodology

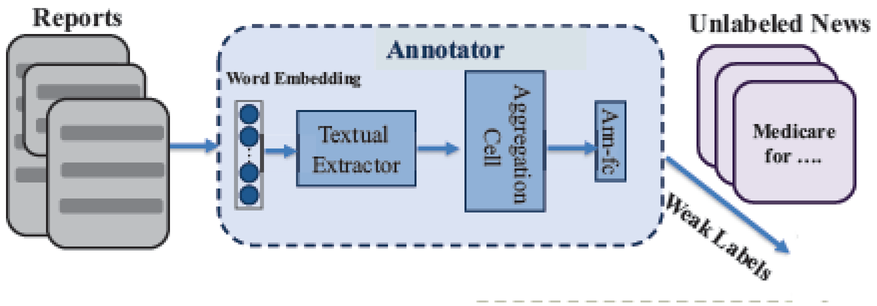
- **<u>Annotator:</u>** pretrained model on reports with their labels that can assign weak labels on unlabeled samples.
- **<u>Reinforced Selector:</u>** Designed to automatically choose high-quality samples, by exploiting reinforcement learning techniques.
- **<u>Fake News Detector:</u>** We use the originally labeled samples and the selected samples from the selector to train the Fake News Detector
- **<u>Textual Feature Extractor:</u>** Both the annotator and the fake news detector modules use a Textual Feature Extractor based on a convolutions neural network



Annotator

Automatic Annotation based on Reports

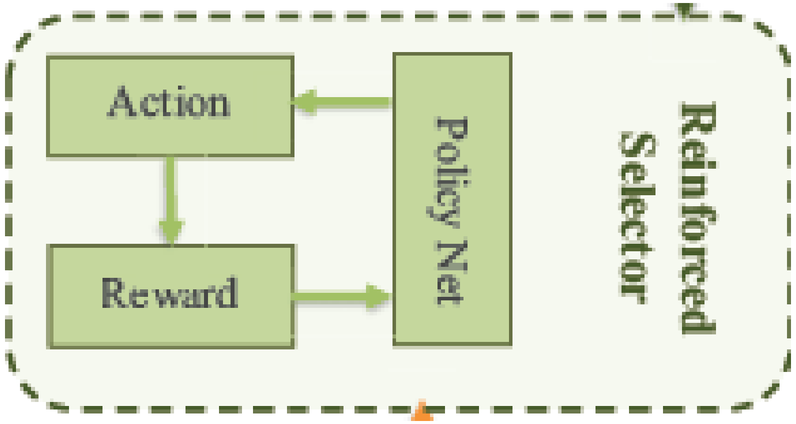
The report messages from multiple users for one piece of news are permutation invariant. Therefore, there can be designed an aggregation cell consisting of a commutative aggregation function and a fully connected layer.



Given the unlabeled news set *X* with corresponding reports as input
we use the trained annotator to predict their labels, which are denoted as Y.

This can produce noisy results and thus the challenge is to select high quality samples from Y set.

Data Selection via Reinforcement Learning 1/2



- The objective of the data selector is to automatically select high-quality samples from those with weak labels obtained from the annotator.

- According to this criteria, we design a performance-driven data selection method (called reinforced data selector) using reinforcement learning mechanism.

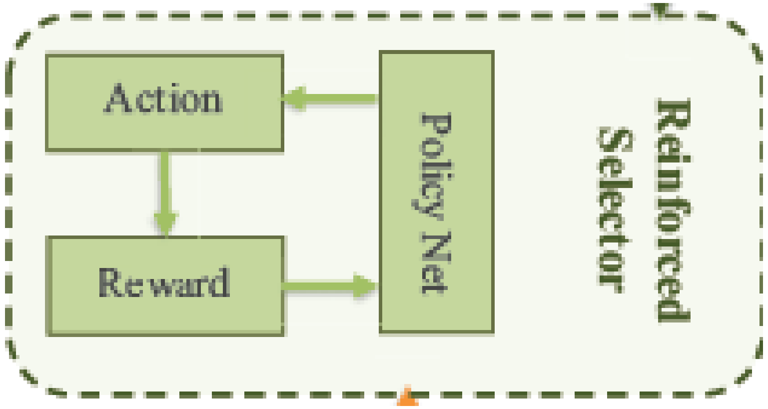
- The selector divides the whole dataset into K small bags of data samples, i.e., $X = \{ X^{(k)} \}_{k=1}^{K}$.

- By using multiple small bags of samples we provide more feedback to the selector and this makes the training procedure of reinforcement learning more efficient

- For every sample, the *action* of reinforced data selector is to retain or remove. The decision of the current sample $x^{(k)}_i$ is based on its *state* vector and all the previous decisions of samples $\{x^{(k)}_1, x^{(k)}_2, ..., x^{(k)}_{i-1}\}$.

Data Selection via Reinforcement Learning 2/2

- State Vector [($S_{(k)i}$]
  - output probability from the annotator
  - the output probability from fake news detector
  - the maximum of cosine similarity between the current sample and the chosen samples
  - the weak label of the current sample
- Action
  - 1 represents the action to *retain* the sample
  - 0 represents the action to *remove* the sample
  - This result is produced by a Policy network which includes two fully connected layers with corresponding activation functions



- Reward
  - The reward of Data Selection via Reinforcement Learning is the maximized accuracy on detecting the fake news on any sample.

Fake News Detector

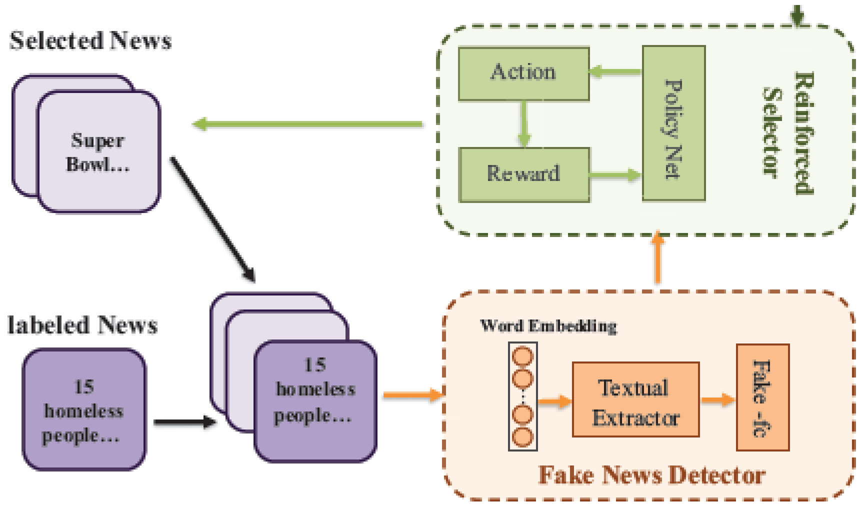
The final component of the article is the fake news detector.

The fake news detector is a neural network, which consists

- textual feature extractor and
- a fully-connected layer, namely Fake-fc, with corresponding activation functions.

The input to fake news detector is news content, and the output is the probability of the given news being fake.

The fake news detector can predict the label of all the news articles by a model trained on the enhanced training set generated by both annotator and reinforced selector.



Experiments

Table 1: The Statistics of the WeChat Datasets.

		# News	# Report	# Avg. Reports/News
Unlabeled	-	22981	31170	1.36
Labeled Training	Fake	1220	2010	1.65
	Real	1220	1740	1.43
Labeled Testing	Fake	870	1640	1.89
	Real	870	1411	1.62

Table 2: The performance comparison of different methods on WeChat dataset.

Category	Method	Accuracy	AUC-ROC	Fake News			Real News		
				Precision	Recall	F_1	Precision	Recall	F_1
Supervised	LIWC-LR	0.528	0.558	0.604	0.160	0.253	0.517	0.896	0.655
	LIWC-SVM	0.568	0.598	0.574	0.521	0.546	0.563	0.614	0.587
	LIWC-RF	0.590	0.616	0.613	0.483	0.541	0.574	0.696	0.629
	LSTM	0.733	0.799	0.876	0.543	0.670	0.669	**0.923**	0.775
	CNN	0.747	0.834	0.869	0.580	0.696	0.685	0.913	0.783
	EANN	0.767	0.803	0.863	0.634	0.731	0.711	0.899	0.794
Semi-supervised	$LSTM_{semi}$	0.753	0.841	0.854	0.611	0.713	0.697	0.895	0.784
	CNN_{semi}	0.759	0.848	0.850	0.630	0.723	0.706	0.889	0.787
Automatically annotated	WeFEND−	0.807	0.858	0.846	**0.751**	0.795	0.776	0.863	0.817
	WeFEND	**0.824**	**0.873**	**0.880**	**0.751**	**0.810**	**0.783**	0.898	**0.836**

Thank you ☺