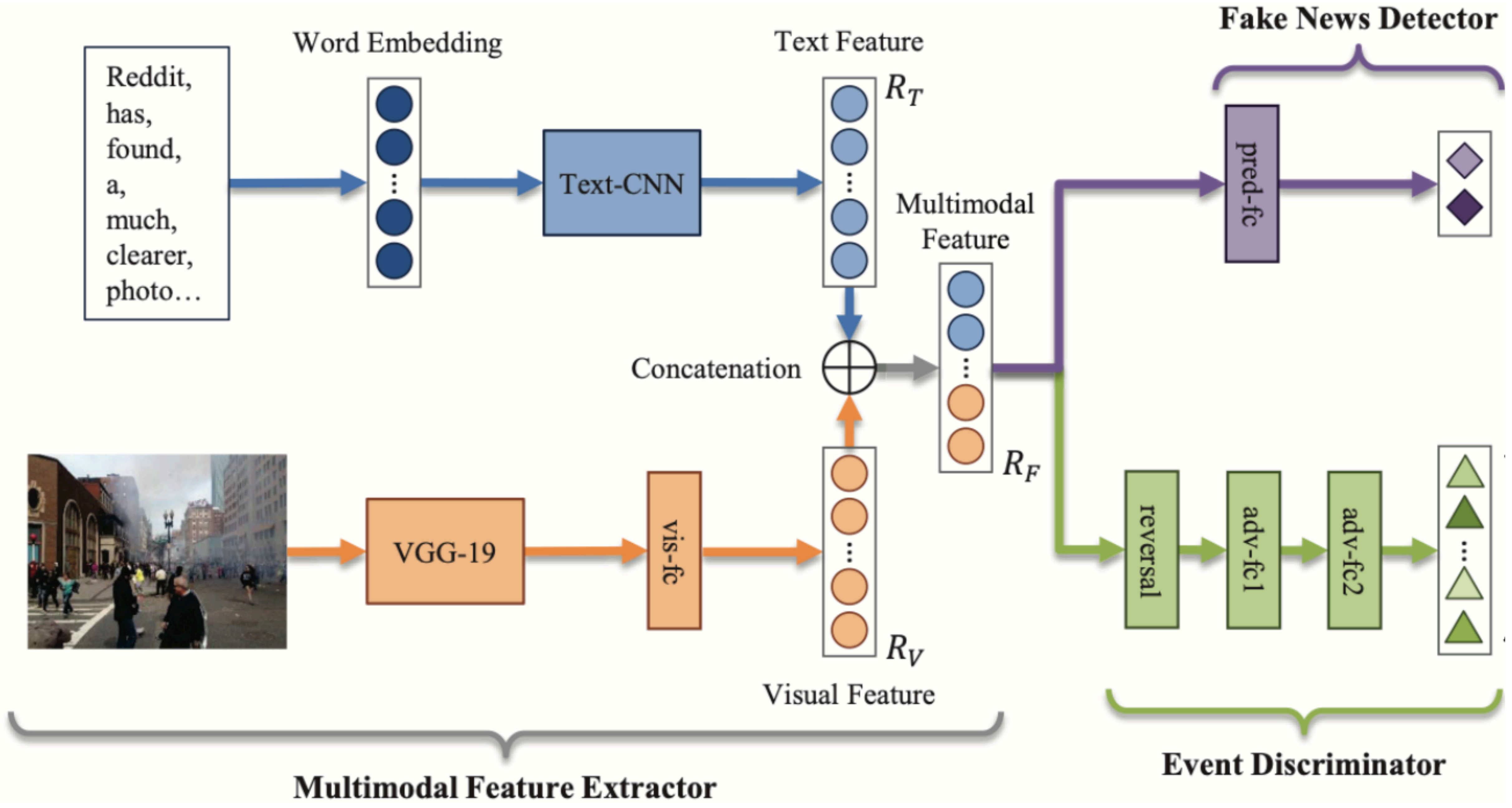
The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Weak Supervision for Fake News Detection via Reinforcement Learning

Yaqing Wang,[1] Weifeng Yang,[2] Fenglong Ma,[3] Jin Xu,[2*] Bin Zhong,[2] Qiang Deng,[2] Jing Gao[1*]
[1]State University of New York at Buffalo, New York, USA
[2]Data Quality Team, WeChat, Tencent Inc., China
[3]Pennsylvania State University, Pennsylvania, USA
[1]{yaqingwa, jing}@buffalo.edu, [2]{curryyang, jinxxu, harryzhong, calvindeng}@tencent.com, [3]fenglong@psu.edu

AAAI'20

210723 Chia-Chun Ho



EANN-KDD18

EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection
Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, Jing Gao

SUNY Buffalo. KDD, 2018.

Dataset

We recently release a dataset (in Chinese) on fake news from Wechat. The dataset includes news titile, report content, news url and image url. Find more details via https://github.com/yaqingwang/WeFEND-AAAI20

The data folder contains a subset of weibo dataset for a quick start. If you are interested in full weibo dataset, you can download it via https://drive.google.com/file/d/14VQ7EWPiFeGzxp3XC2DeEHi-BEisDINn/view?usp=sharing. (Approximately 1.3GB)

Main Idea

One of the unique challenges for fake news detection on social media is how to identify fake news on **newly emerged events.** The EANN is desgined to **extract shared features among all events** to effectively improve the performance of fake news detection on never-seen events.

Outline

Introduction

Methodology

Experiments

Conclusions

Comments

Introduction

Fake News Detection

- Roughly divided fake news detection into two categories:
    - Traditional learning methods
        - Extract features from news articles and train classifiers based on the extracted features
    - Deep learning models
        - Learning informative representations automatically
        - Usually require a large amount of hand-labeled data (i.e. $\hat{y} = 1$ or 0)

Introduction

Labeling fake news

- Creations of such data is expensive and time-consuming
- Accurate labels can only be obtained when the annotators have sufficient knowledge about the events.
- Dynamic nature of news articles leads to decaying quality of existing labeled samples.
    - Some of these samples may become outdated quickly and can't represent the news articles on newly emerged events.
    - Annotators have to continuously label newly emerging news articles, which is infeasible
        - It's essential to tackle the challenge of labeling fake news

Introduction

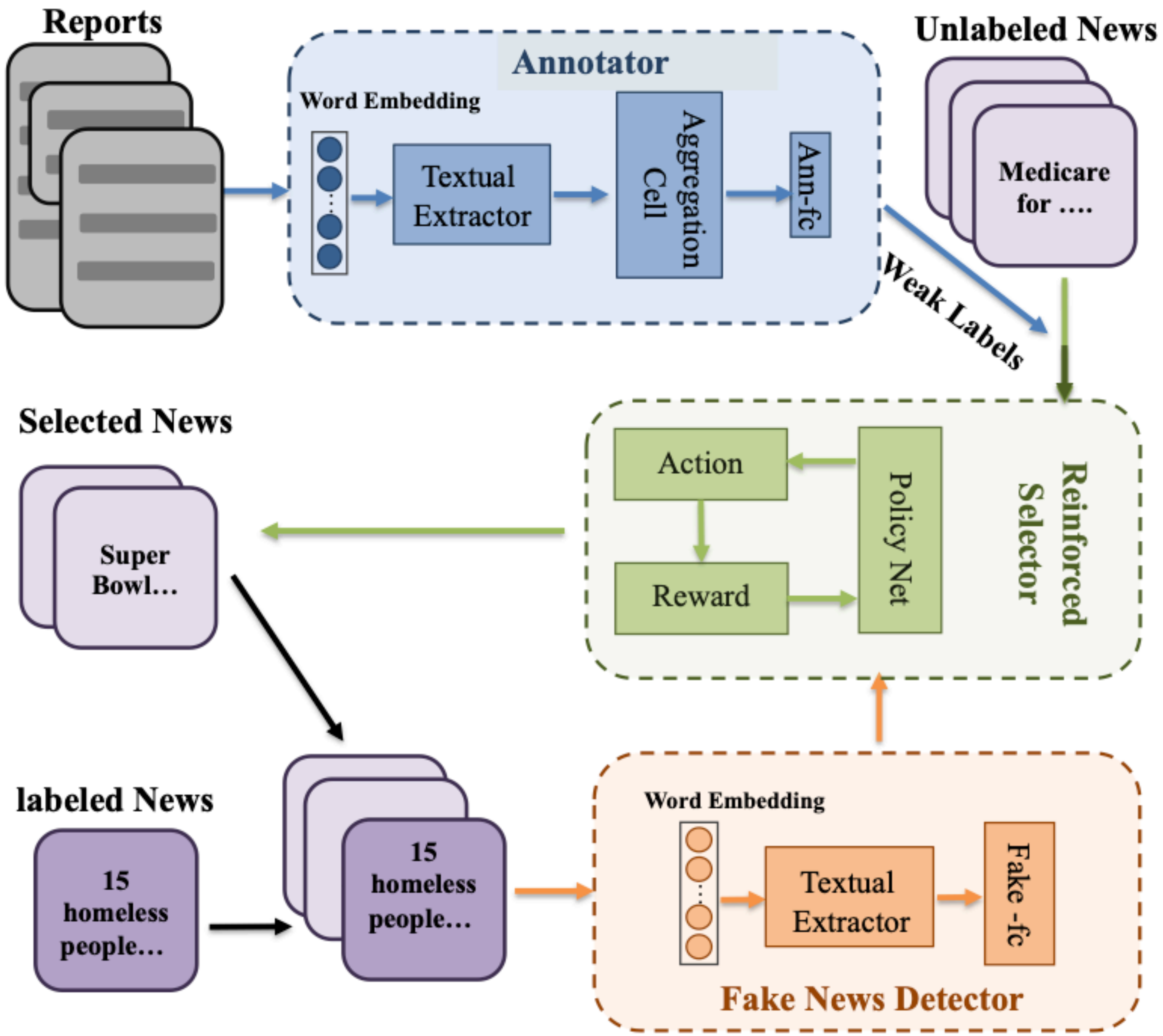
Leverage the feedback provided by users who read the news

- A news article published on a WeChat official account, a user who reads the article can report whether this news is fake or not with a brief explanation.
- Such reports from users can be regarded as "weak" annotation for the task of fake news detection
  - The large collection of user reports can help alleviate the label shortage problem in fake news detection
  - These weak annotated samples are unavoidably compared with expert-labeled samples
    - Users may report real news as fake or the reasons they provide may not be meaningful
    - <u>Transform weak annotation to labeled samples in the training set and select high-quality samples</u> is the major issue need to solve

Introduction

Reinforced Weakly-supervised FakE News Detection framework (WeFEND)

- Leverage the crowd users' feedback as weak supervision for fake news detection
- Consists of three main components:
  - Annotator 🟦
  - Reinforced selector 🟩
  - Fake news detector 🟧



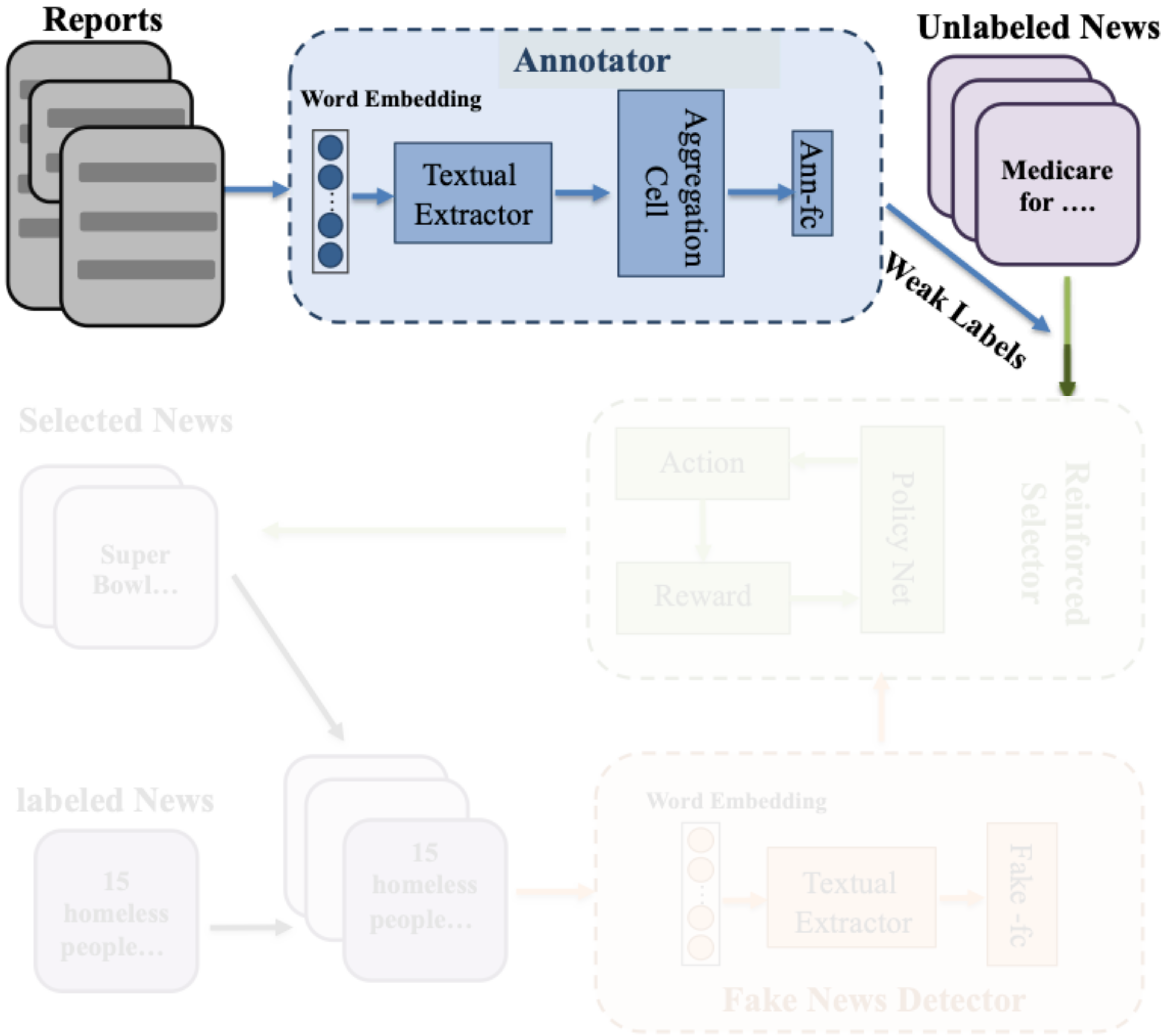
Introduction

WeFEND: annotator

- Given a small set of labeled fake news samples together with users' feedback towards these news articles

- Annotator can be seen as a pretrained model on the reports with their labels

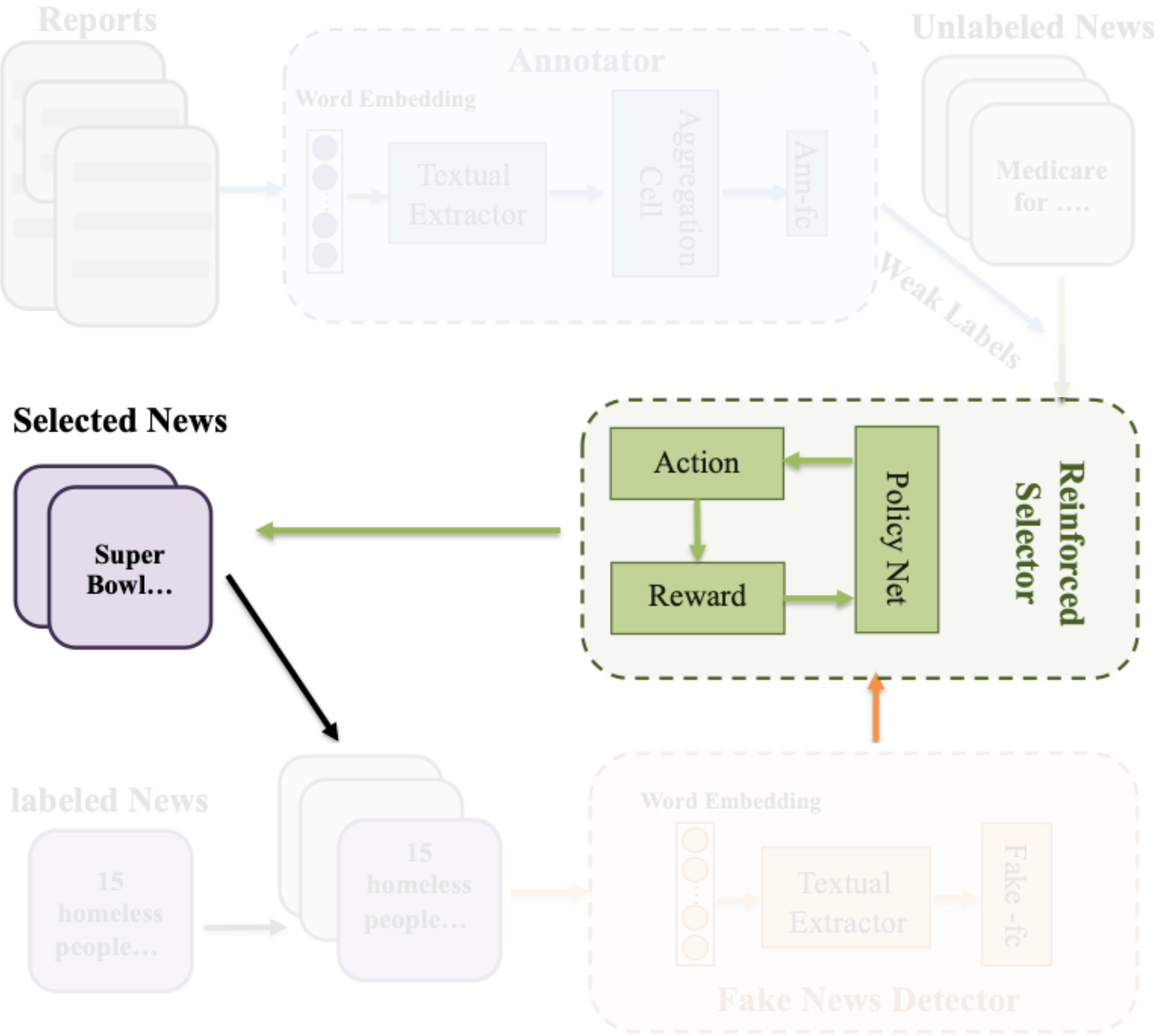
- Train an annotator based on the feedback can automatically assign weak labels for those unlabeled news articles simply based on the user feedback they received



Introduction

WeFEND: reinforced selector

- It's hard to guarantee the quality of weak labels
- The reinforced selector which employs **reinforcement learning** techniques then selects high-quality samples from weakly labeled samples as input to the fake news detector.



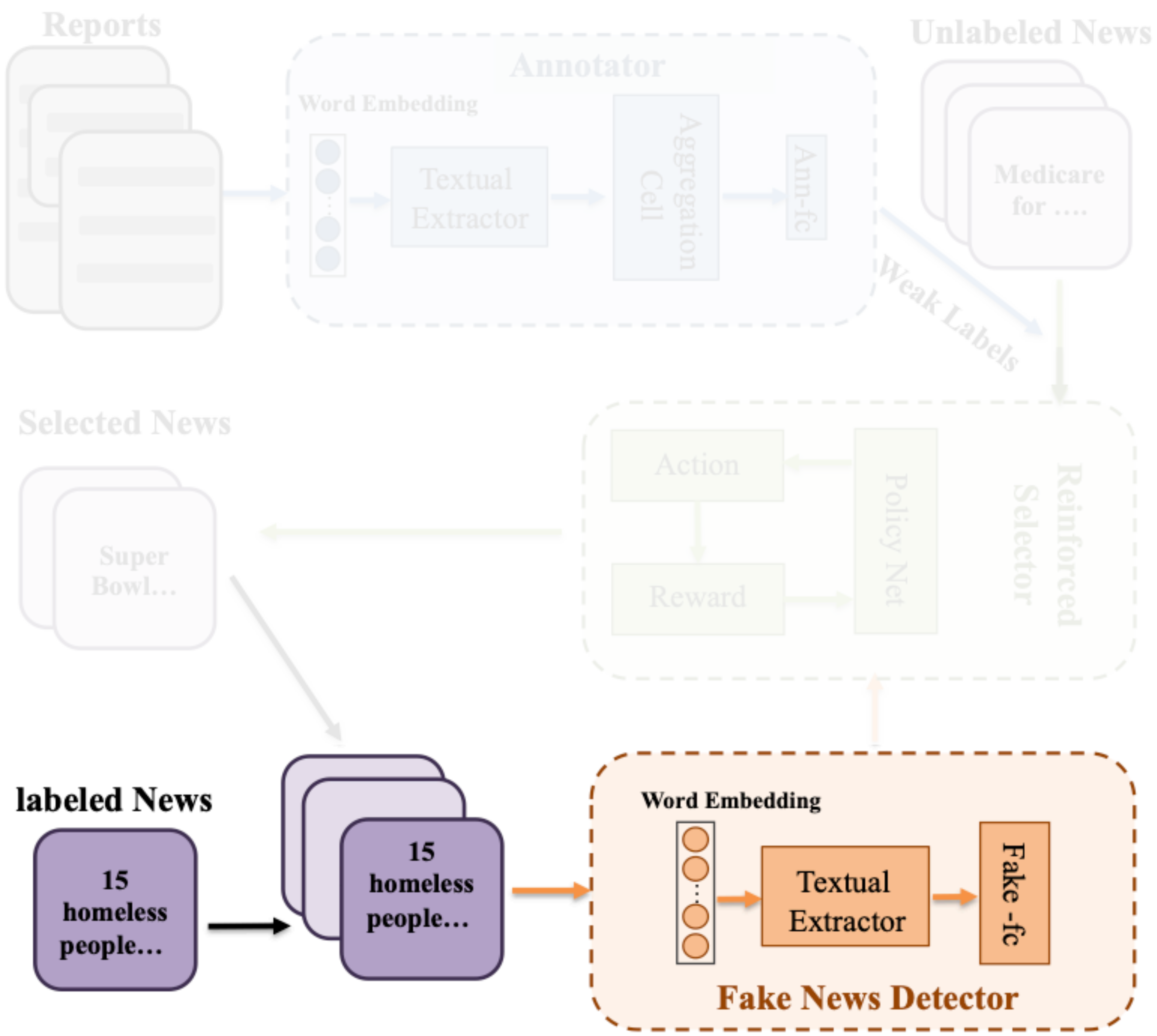
Introduction

WeFEND: fake news detector

- The selected samples and the original labeled samples are used to train fake news detector

- The fake news detector finally assigns a label for each input article based on its content

- The three components integrate nicely and their interactions mutually enhance their performance



Introduction

Contributions

- Recognize the label shortage issue and proposed to leverage user reports as weak supervision for fake news detection
- WeFEND framework can automatically annotate news articles, which help enlarge the size of the training set to ensure the success in deep learning models
- Adopting reinforcement learning techniques, WeFEND has the ability of selecting high-quality samples, which further leads to the improvement of the fake news detection performance

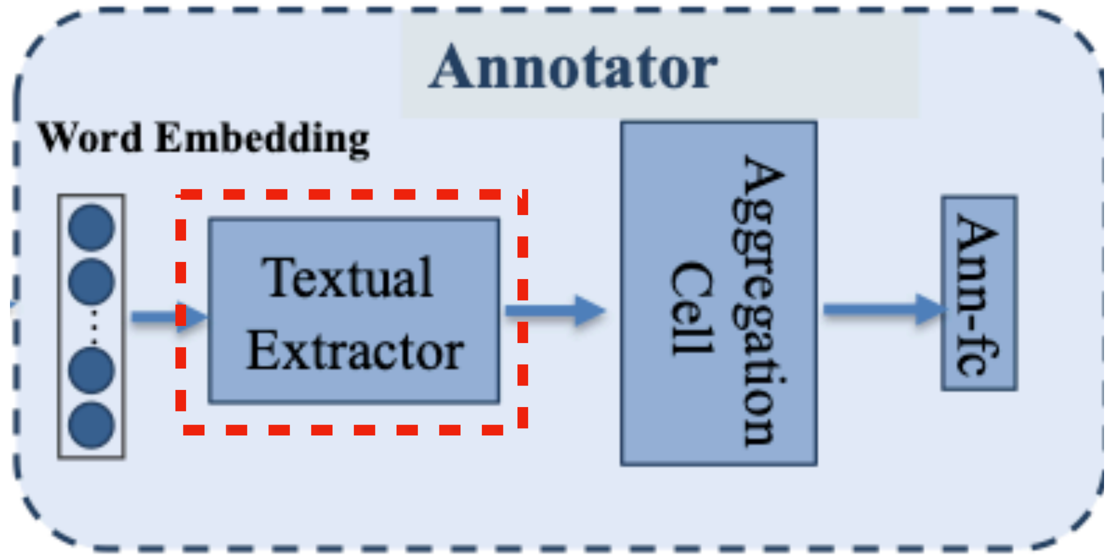
Methodology

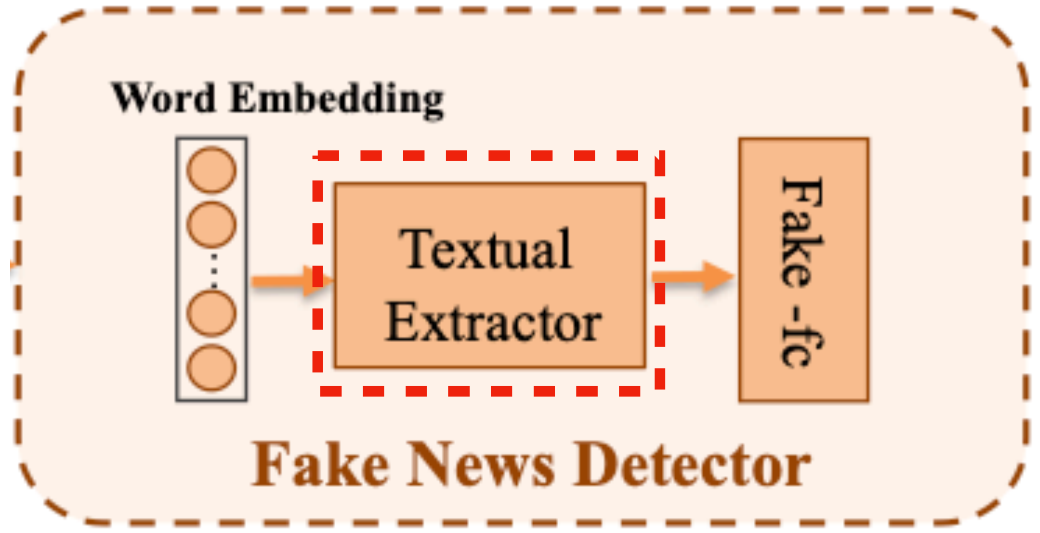
Overview

- Each sample consists of both news articles and user feedback comments (*reports*)
    - Both are texts, and are transformed into vector representations by word embedding
    - User feedback comments are detailed reasons and evidence provided by user
- A small set of samples are labeled by experts as fake or real
- Our objective is to predict the labels of the unlabeled samples

Methodology

Textual Feature Extractor

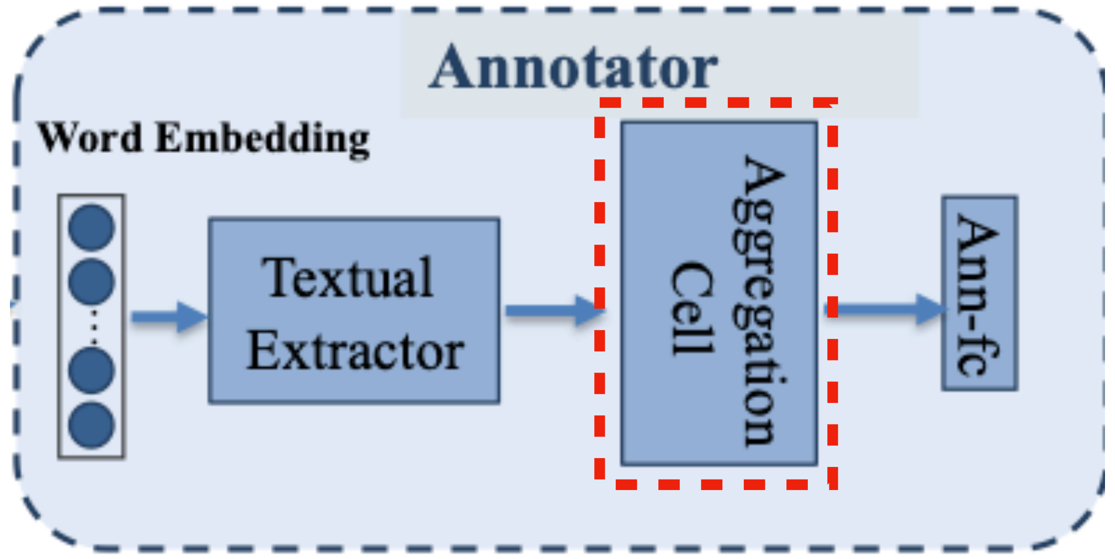




- Use Text-CNN as textual feature extractor
- The input of the textual feature extractor is
  - news content (in fake news detector) or
  - a report message (in annotator)
- The learned representation from textual feature extractor are the input features to annotator and fake news detector

Methodology

Automatic Annotation based on Reports



- One news article may have reports from multiple users

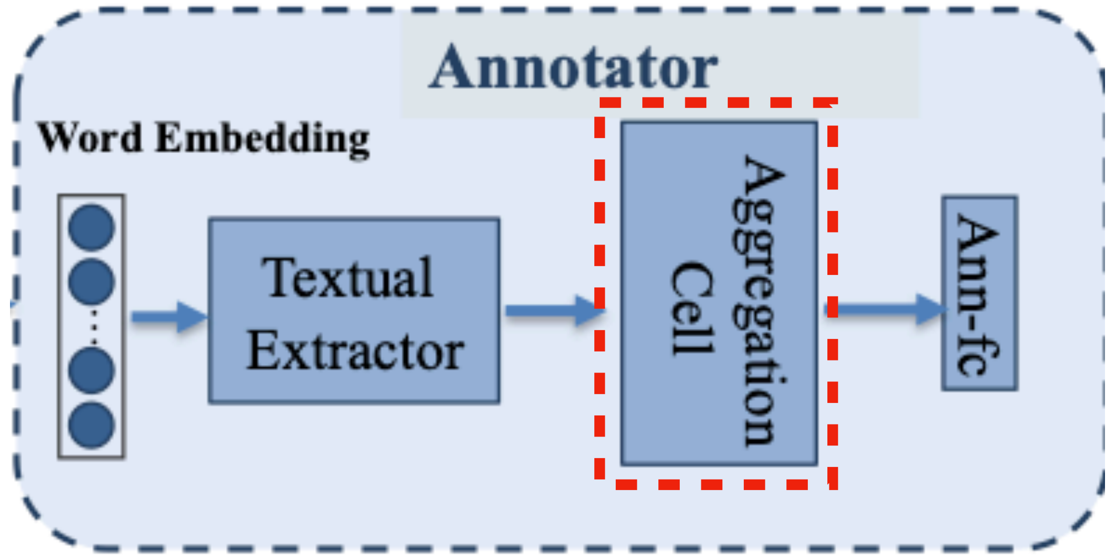
- Propose to aggregate features obtained from different reports for one sample.

- Design an aggregation cell consisting of a commutative aggregation function and a fully-connected layer

- The commutative aggregation function, like sum, mean and max-pooling, can combine the permutation invariant input set

Methodology

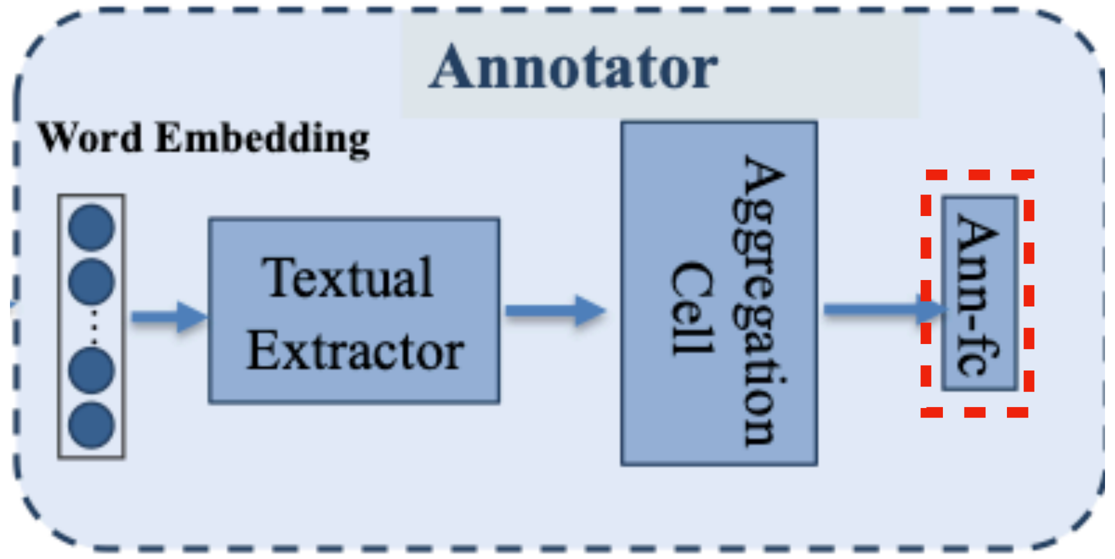
Automatic Annotation based on Reports



- Take the i-th sample as an example, and the j-th report message: $r_j^{(i)}$
- The corresponding report message set is denoted as $R^{(i)} = \{r_1^{(i)}, r_2^{(i)}, \cdots, r_{\left|R^{(i)}\right|}^{(i)}\}$
    - $\left|R^{(i)}\right|$: number of report messages of i-th sample
- $r_j^{(i)} \in R^{(i)}$ is first fed into the textual feature extractor to obtain $\mathbf{h}_j^{(i)}$
- Use the aggregation cell to combine $R^{(i)}$ to learn the hidden feature representation $\mathbf{h}^{(i)}$
- Procedure of aggregation cell: $\mathbf{h}^{(i)} = \mathrm{ReLU}\left(\mathbf{w}_r \cdot \sum_{j=1}^{\left|R^{(i)}\right|} \frac{\mathbf{h}_j^{(i)}}{\left|R^{(i)}\right|}\right)$, $\mathbf{w}_r$: weight of the fully-connected layer

Methodology

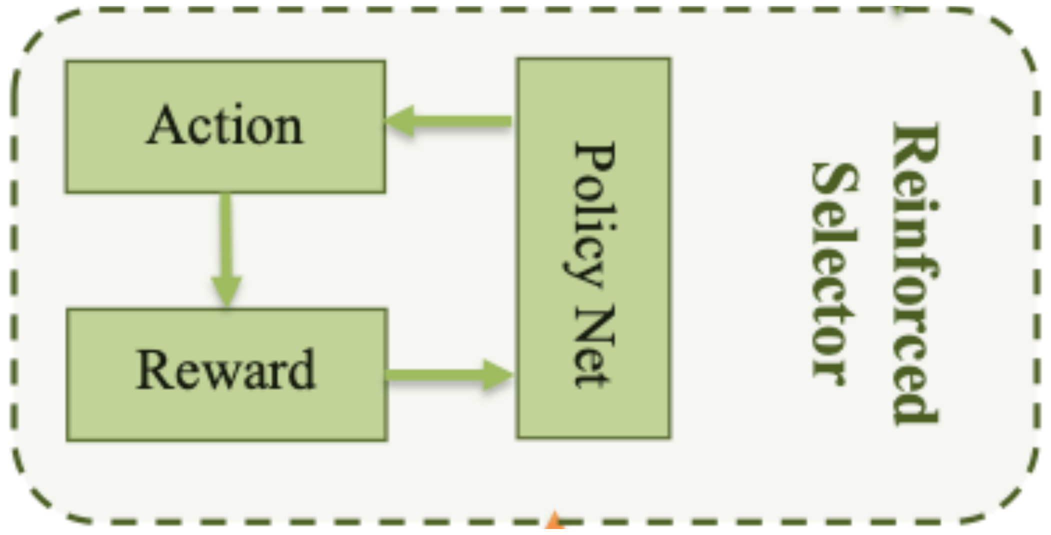
Automatic Annotation based on Reports



- Feed $\mathbf{h}^{(i)}$ into the fully connected layer, denoted as Ann-fc, to output the corresponding probability of the i-th sample being a fake one
    - $D_r\left(R^{(i)};\theta_r\right)$, θ_r: all parameters of the annotator and corresponding textual feature extractor
- Entire report message dataset $R = \{R^{(1)}, R^{(2)}, \cdots, R^{(\left|R\right|)}\}$, $\left|R\right|$: number of report sets
- Corresponding ground truth labels of news $Y = \{y^{(1)}, y^{(2)}, \cdots, y^{(\left|R\right|)}\}$
- Loss function for the proposed annotator is defined by cross entropy as follows:
    - $$L_r\left(R, Y;\theta_r\right) = -\frac{1}{|R|}\sum_{i=1}^{|R|}\left[y^{(i)}\log D_r\left(R^{(i)};\theta_r\right) + \left(1 - y^{(i)}\right)\log\left(1 - D_r\left(R^{(i)};\theta_r\right)\right)\right]$$

Methodology

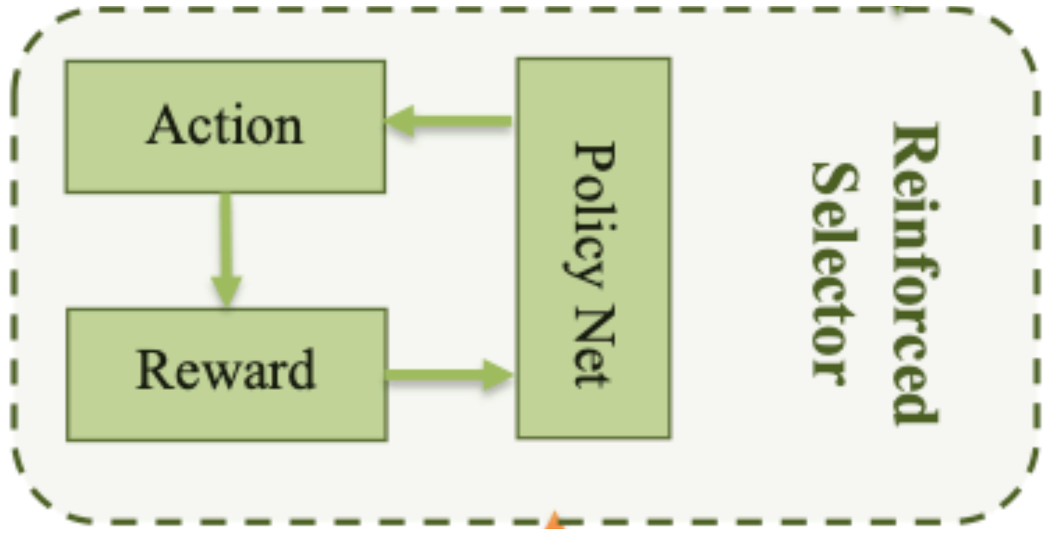
Data Selection via Reinforcement Learning



- The criteria of the selection is based on whether adding the chosen sample cam improve the fake news detection performance
  - Design a performance-driven data selection method using reinforcement learning mechanism.
- $\tilde{X}$: all the input data of the proposed reinforced data selector
- Instead of directly putting the entire dataset $\tilde{X}$ into the selector, divide $\tilde{X}$ into K small bags of data examples: $\tilde{X} = \left\{\tilde{X}^{(k)}\right\}_{k=1}^{K}$
- For the k-th bag of data contains B samples: $\tilde{X}^{(k)} = \{x_1^{(k)}, x_2^{(k)}, \cdots, x_B^{(k)}\}$
- Using multiple small bags of samples can provide more feedback to selector and makes the training procedure of reinforcement learning more efficient

Methodology

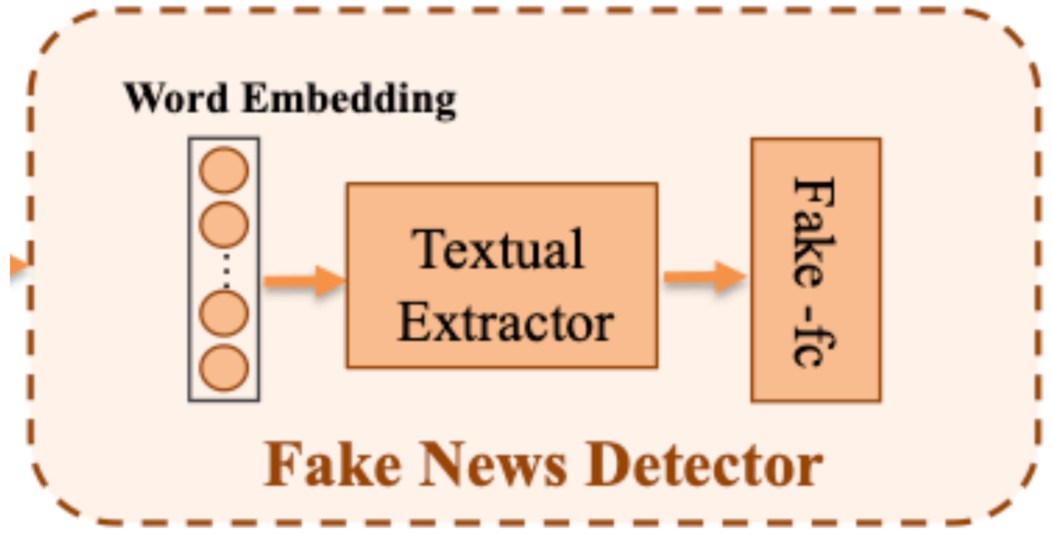
Data Selection via Reinforcement Learning



- For every sample, the *action* of reinforced data selector is to *retain* or *remove*.
- The decision of the current sample $x_i^{(k)}$ is based on its *state* vector and all previous decisions of samples $\{x_1^{(k)}, x_2^{(k)}, \cdots, x_{i-1}^{(k)}\}$
- The data selection problem can be naturally cast as a Markov Decision Process (MDP)
- Since the goal of data selection is to improve the performance of fake news detection, directly use the performance (accuracy) changes of fake news detection as the *reward* for reinforced selector

Methodology

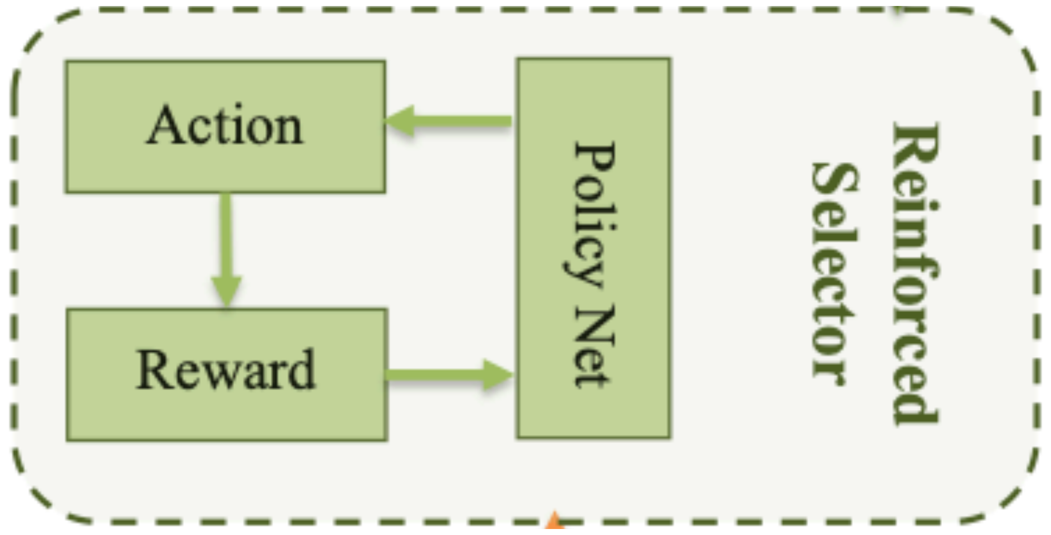
Fake news detector



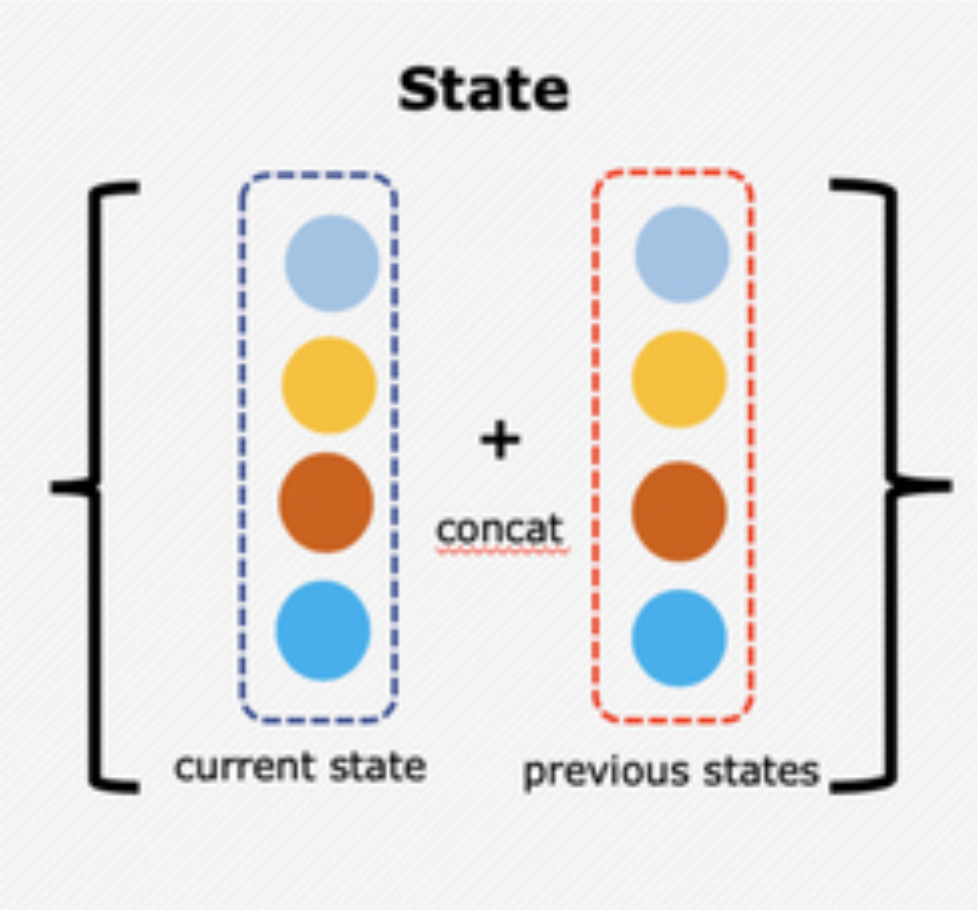
- Consists of a textual feature extractor and a fully-connected layer, namely Fake-fc
- Input: news content
- Output: the probability of the given news being fake
- $D_n(\cdot\,;\theta_n)$, θ_n: all the parameters

Methodology

Data Selection via Reinforcement Learning - *State*

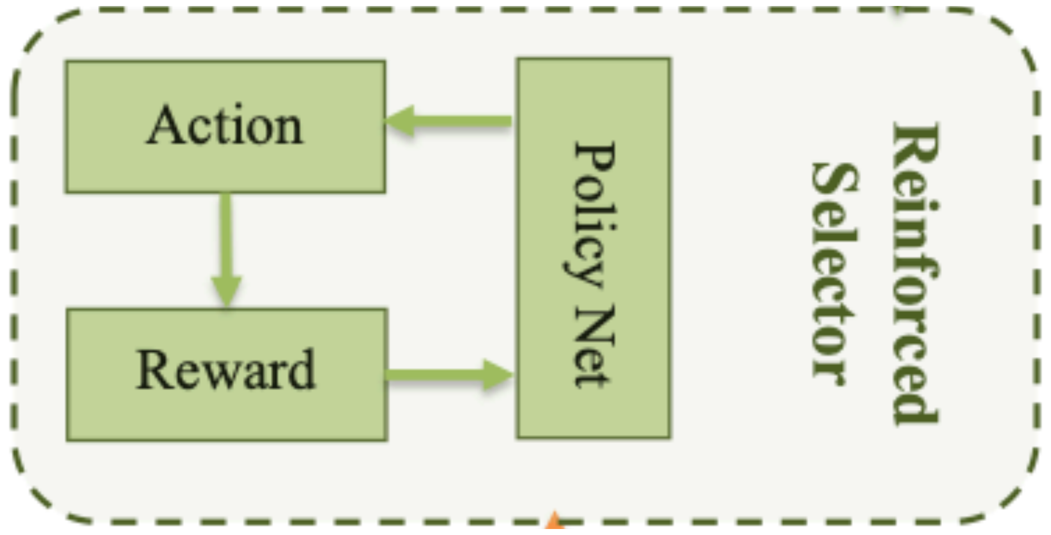


- $s_i^{(k)}$: state vector of the sample $x_i^{(k)}$
- Every action is made based on the current sample and the chosen sample, the state vector mainly consists of two components:
    - Representation of the current sample (related to data quality and diversity)
    - Average representation of the chosen samples
- The concatenation of the current state vector and the average of previous state vectors is considered as the final state vector $s_i^{(k)}$



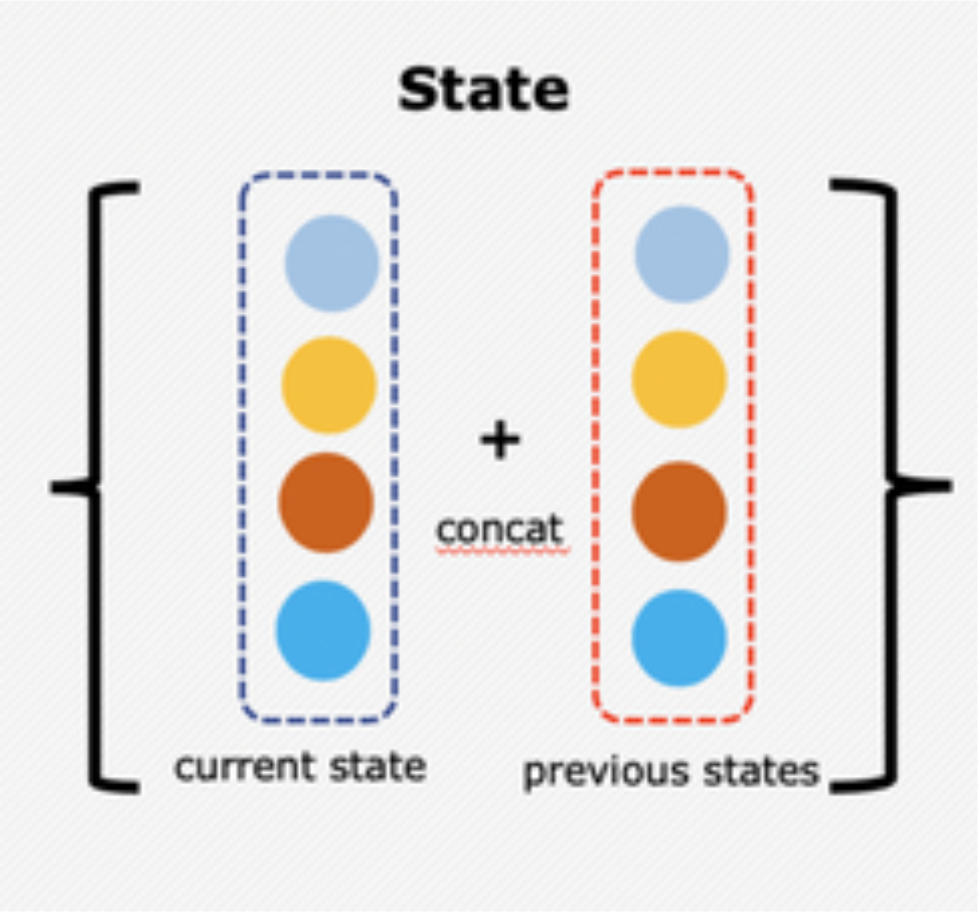
Methodology

Data Selection via Reinforcement Learning - *State*



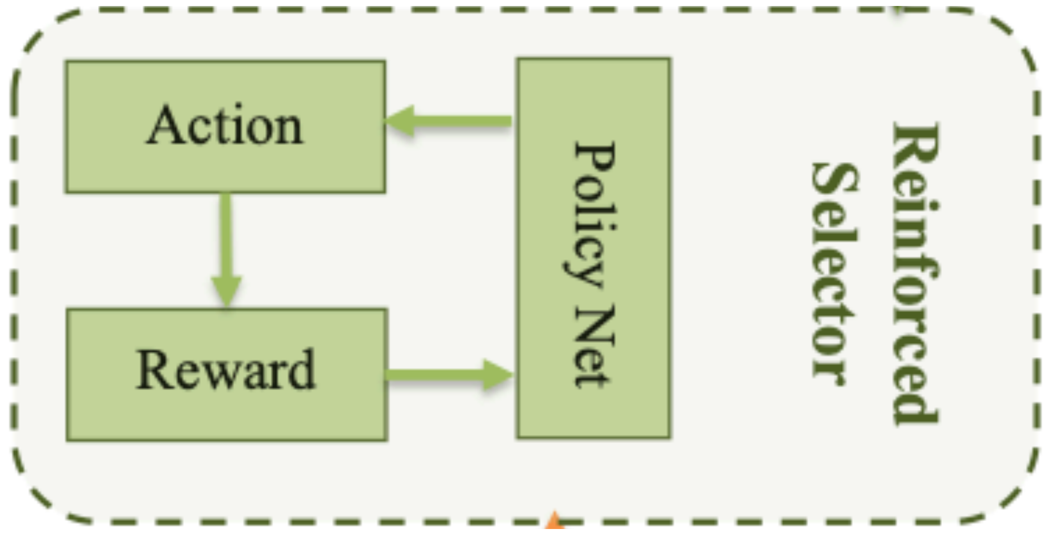
- The current state vector contains four elements:

    1) output probability from the annotator (quality) 🟦
    2) output probability from fake news detector (quality) 🟨
    3) maximum of cosine similarity between the current sample and the chosen samples (diversity) 🟧
    4) weak label of the current sample (balance the distribution of classes) 🟦



Methodology

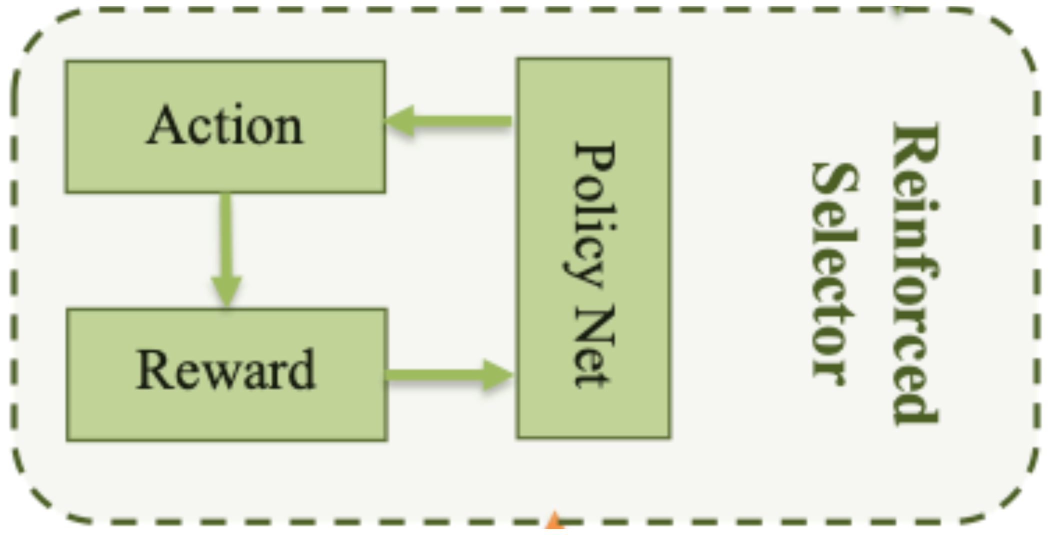
Data Selection via Reinforcement Learning - *Action*



- The action value $a_i^{(k)}$ for every sample is 1 or 0.
    - 1: the action to *retain* the sample
    - 0: the action to *remove* the sample
- To determine the action, train a policy network includes two fully connected layers with corresponding activation functions, denote as $P\left(\,\cdot\,;\theta_s\right)$, θ_s: the parameters
    - $P\left(s_i^{(k)};\theta_s\right) = \delta\left(\mathbf{w}_{s2}\cdot\mathrm{ReLU}\left(\mathbf{w}_{s1}\cdot s_i^{(k)}\right)\right)$
    - $\mathbf{w}_{s1}, \mathbf{w}_{s2}$: weights of fully-connected layer, δ: sigmoid activation function

Methodology

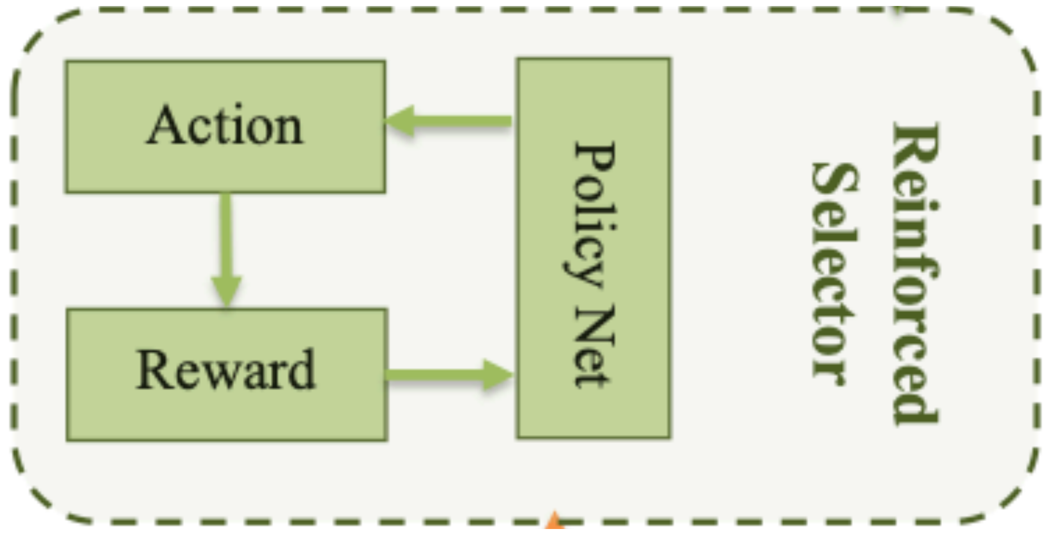
Data Selection via Reinforcement Learning - *Action*



- Then the action $a_i^{(k)}$ is sampled according to the output probability.
- The policy can be represented as:
- $\pi_{\theta_s}\left(s_i^{(k)}, a_i^{(k)}\right) = \begin{cases} p_i^{(k)} & \text{if } a_i^{(k)} = 1 \\ 1 - p_i^{(k)} & \text{if } a_i^{(k)} = 0 \end{cases}$

Methodology

Data Selection via Reinforcement Learning - *Reward*



- Use performance changes of detection model $D_n\left(\cdot\,;\theta_n\right)$ as the reward function

- Given $\tilde{X}^{(k)} = \{x_1^{(k)}, x_2^{(k)}, \cdots, x_B^{(k)}\}$, the actions of retaining or removing are made based on the probability output from the policy network

  - To evaluate the performance changes, need to set a baseline accuracy acc

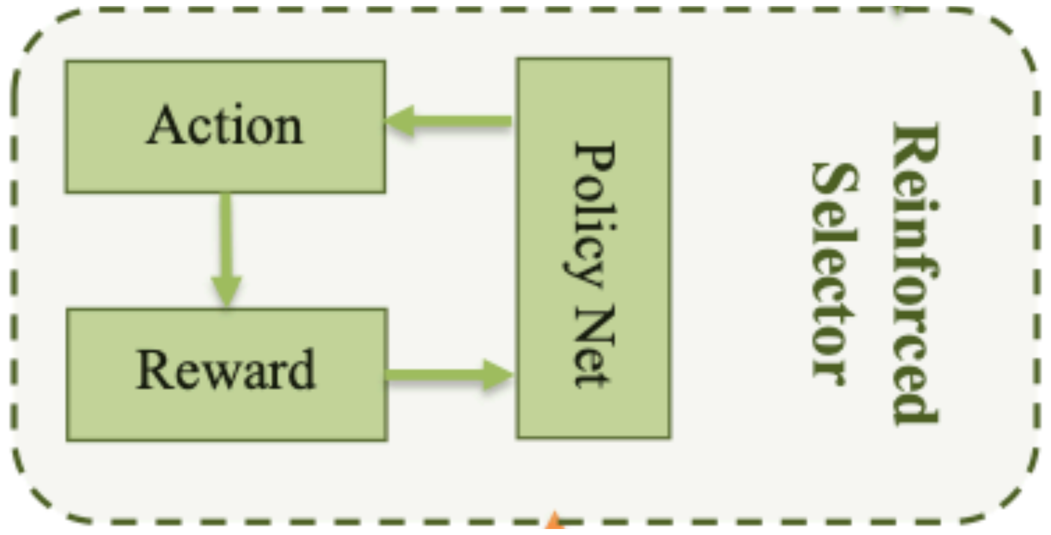
  - Calculate acc with $D_n\left(\cdot\,;\theta_n\right)$ on validation dataset

  - Then new accuracy acc_k can obtained with the retrained model

- Finally, the reward R_k for k-th bag data $\left\{x_i^{(k)}\right\}_{i=1}^{B}$: $R_k = acc_k - acc$

Methodology

Data Selection via Reinforcement Learning - *Reward*



- For k-th bag data $\left\{x_i^{(k)}\right\}_{i=1}^{B}$, aim to maximize the expected total reward

- Since the scale of R_k is small use the summation of reward to define the objective function:

- $J\left(\theta_s\right) = \sum_{i=1}^{B} \pi_{\theta_s}\left(s_i^{(k)}, a_i^{(k)}\right) R_k$

Methodology

Reinforced Weakly-supervised Fake News Detection Framework

- First, pre-train the annotator using the labeled report data $\{R, Y\}$ and assign weak labels $\hat{Y}^u$ to the unlabeled news set X^u
- The proposed reinforced selector will select high-quality samples $\{X_s, Y_s\} = \left\{X_s^{(k)}, Y_s^{(k)}\right\}_{k=1}^{K}$ from the weakly labeled dataset $\{X^u, \hat{Y}^u\}$
- Then both selected dataset $\{X_s, Y_s\}$ and original labeled data $\{X, Y\}$ are fed into the fake news detector for training.

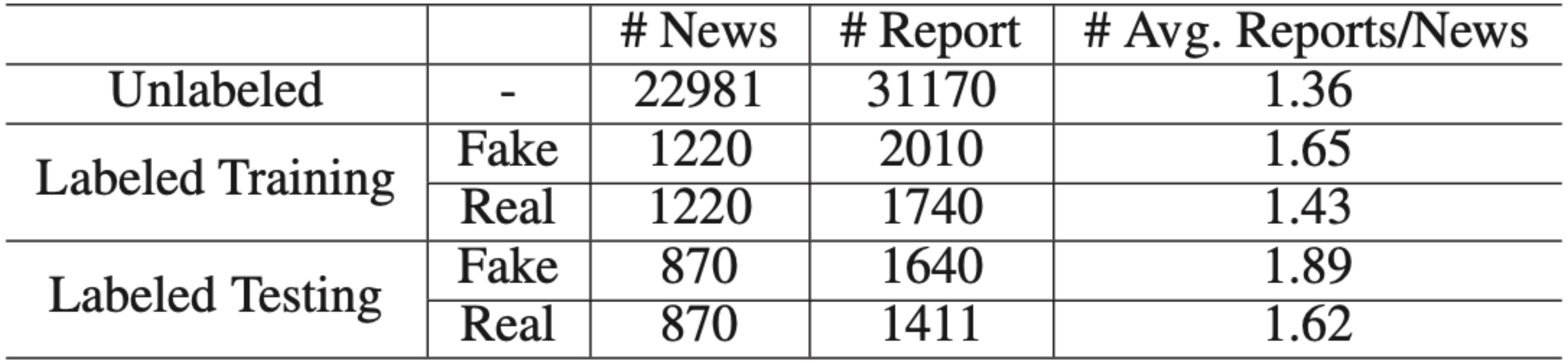
Methodology

Reinforced Weakly-supervised Fake News Detection Framework

- The final loss of fake news detection consists of two sub losses:
- $L_n\left(X, Y, X_s, Y_s; \theta_n\right) = \lambda_l \cdot L_n^l\left(X, Y; \theta_n\right) + \lambda_s \cdot L_n^s\left(X_s, Y_s; \theta_n\right)$
- Simply set the values of λ_l and λ_u as 1
  - Loss on a small amount of manually labeled data:
    - $L_n^l\left(X, Y; \theta_n\right) = -\mathbb{E}_{(x,y)\sim(X,Y)}\left[y\log\left(D_n\left(x; \theta_n\right)\right) + (1-y)\log\left(1 - D_n\left(x; \theta_n\right)\right)\right]$
  - Loss on automatically annotated data set
    - $L_n^s\left(X_s, Y_s; \theta_n\right) = -\mathbb{E}_{\left(x_s, y_s\right)\sim\left(X_s, Y_s\right)}\left[y_s\log\left(D_n\left(x_s, \theta_n\right)\right) + \left(1 - y_s\right)\log\left(1 - D_n\left(x_s; \theta_n\right)\right)\right]$

Experiments

Dataset

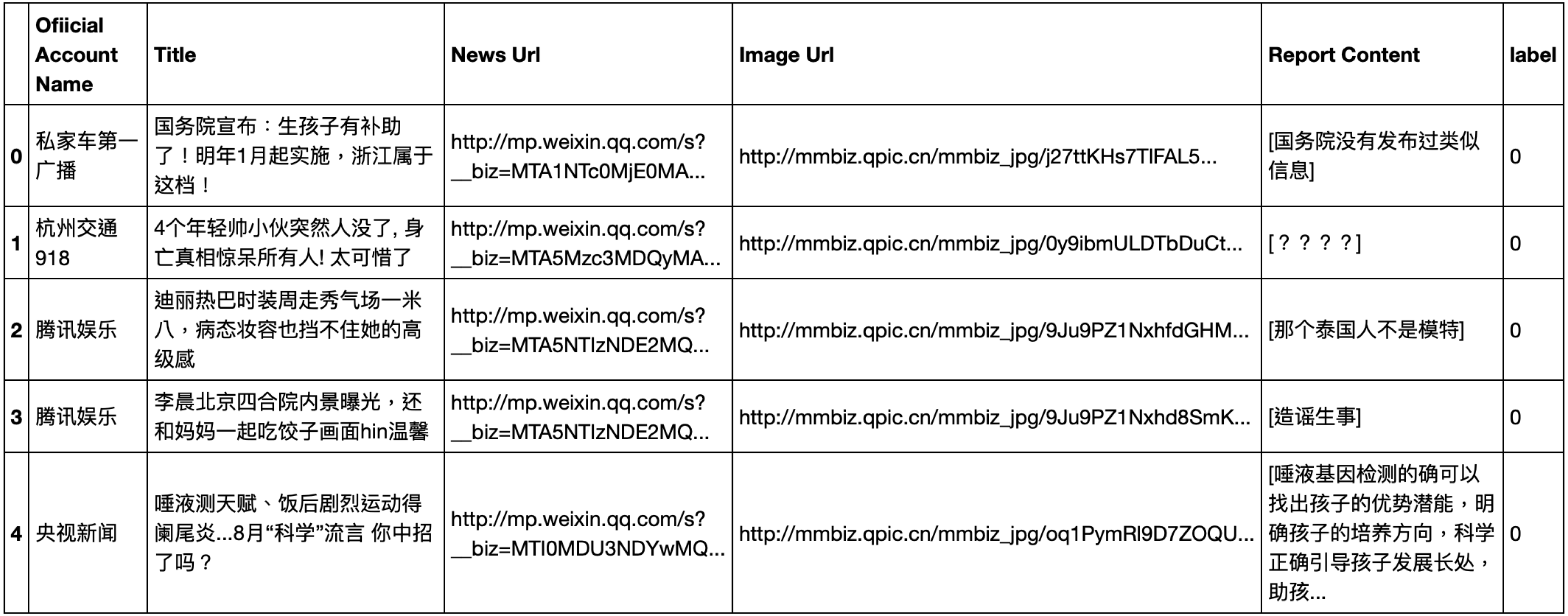


		# News	# Report	# Avg. Reports/News
Unlabeled	-	22981	31170	1.36
Labeled Training	Fake	1220	2010	1.65
	Real	1220	1740	1.43
Labeled Testing	Fake	870	1640	1.89
	Real	870	1411	1.62

- Experiments are conducted on WeChat's Official Accounts
- In this dataset, the news are collected from WeChat's Official Accounts (2018.03-2018.10)
- Split the fake news and real news into training and testing sets according to the post timestamp, training dataset (2018.03-2018.09), testing dataset (2018.09-2018.10)
    - There is no overlapped timestamp of news between these two sets
    - This design is to evaluate the performance of fake news detection on the fresh news
- Also have an unlabeled set containing a large amount of collected news without annotation (2018.09-2018.10)
- Note that the headlines can be seen as the summary of the news content. In the manual annotation process, experts only look at headlines to conduct labeling. Thus, in this paper, use headlines as the input data.

Experiments

Dataset examples



	Ofiicial Account Name	Title	News Url	Image Url	Report Content	label
0	私家车第一广播	国务院宣布：生孩子有补助了！明年1月起实施，浙江属于这档！	http://mp.weixin.qq.com/s?__biz=MTA1NTc0MjE0MA...	http://mmbiz.qpic.cn/mmbiz_jpg/j27ttKHs7TIFAL5...	[国务院没有发布过类似信息]	0
1	杭州交通918	4个年轻帅小伙突然人没了, 身亡真相惊呆所有人! 太可惜了	http://mp.weixin.qq.com/s?__biz=MTA5Mzc3MDQyMA...	http://mmbiz.qpic.cn/mmbiz_jpg/0y9ibmULDTbDuCt...	[？？？？]	0
2	腾讯娱乐	迪丽热巴时装周走秀气场一米八，病态妆容也挡不住她的高级感	http://mp.weixin.qq.com/s?__biz=MTA5NTIzNDE2MQ...	http://mmbiz.qpic.cn/mmbiz_jpg/9Ju9PZ1NxhfdGHM...	[那个泰国人不是模特]	0
3	腾讯娱乐	李晨北京四合院内景曝光，还和妈妈一起吃饺子画面hin温馨	http://mp.weixin.qq.com/s?__biz=MTA5NTIzNDE2MQ...	http://mmbiz.qpic.cn/mmbiz_jpg/9Ju9PZ1Nxhd8SmK...	[造谣生事]	0
4	央视新闻	唾液测天赋、饭后剧烈运动得阑尾炎...8月“科学”流言 你中招了吗？	http://mp.weixin.qq.com/s?__biz=MTI0MDU3NDYwMQ...	http://mmbiz.qpic.cn/mmbiz_jpg/oq1PymRI9D7ZOQU...	[唾液基因检测的确可以找出孩子的优势潜能，明确孩子的培养方向，科学正确引导孩子发展长处，助孩...	0

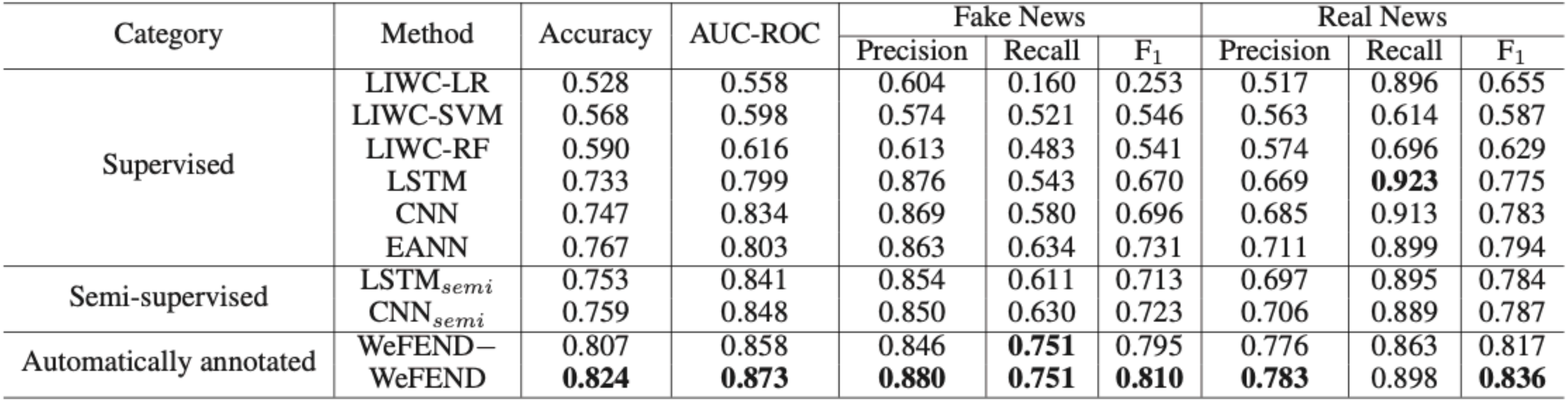
Experiments

Baselines

- **LIWC** (traditional): Logistic Regression (**LIWC-LR**), SVM (**LIWC-SVM**) and Random Forest (**LIWC-RF**)
- **LSTM, CNN (LSTM$_{semi}$, CNN$_{semi}$)**
  - To show effects of automatic annotation, also proposed two semi-supervised models
- **EANN** (KDD'18)
- To show the role of data selector, design one variant of the proposed model named **WeFEND−** , which does not include data selector

Experiments

Performance Comparison

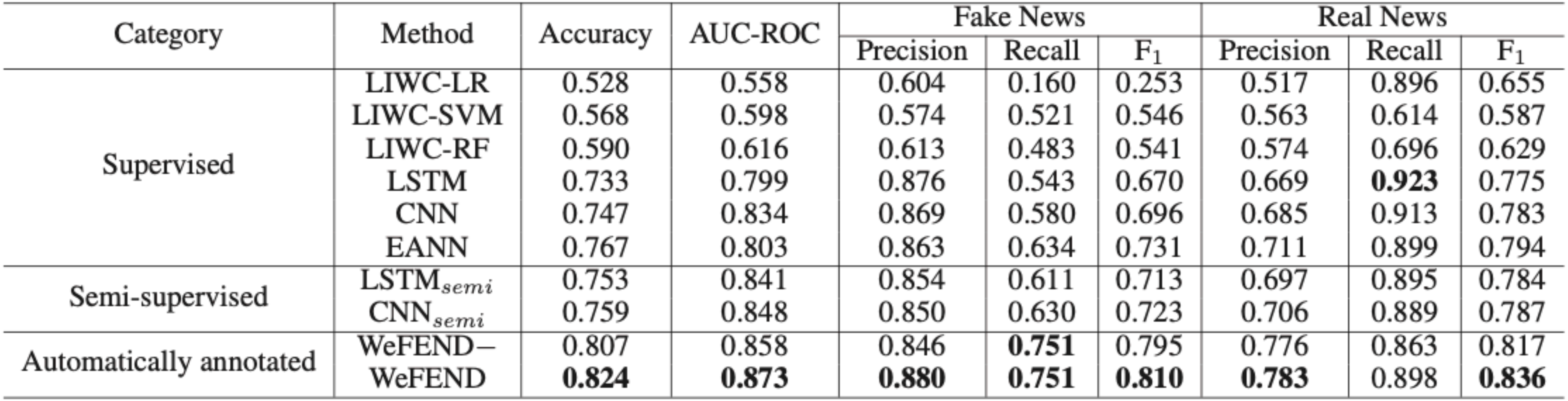


Category	Method	Accuracy	AUC-ROC	Fake News			Real News		
				Precision	Recall	F$_1$	Precision	Recall	F$_1$
Supervised	LIWC-LR	0.528	0.558	0.604	0.160	0.253	0.517	0.896	0.655
	LIWC-SVM	0.568	0.598	0.574	0.521	0.546	0.563	0.614	0.587
	LIWC-RF	0.590	0.616	0.613	0.483	0.541	0.574	0.696	0.629
	LSTM	0.733	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**

- Observe that WeFEND achieves the best results in terms of Accuracy, AUC-ROC, precision, recall and F1 measurement

Experiments

Performance Comparison

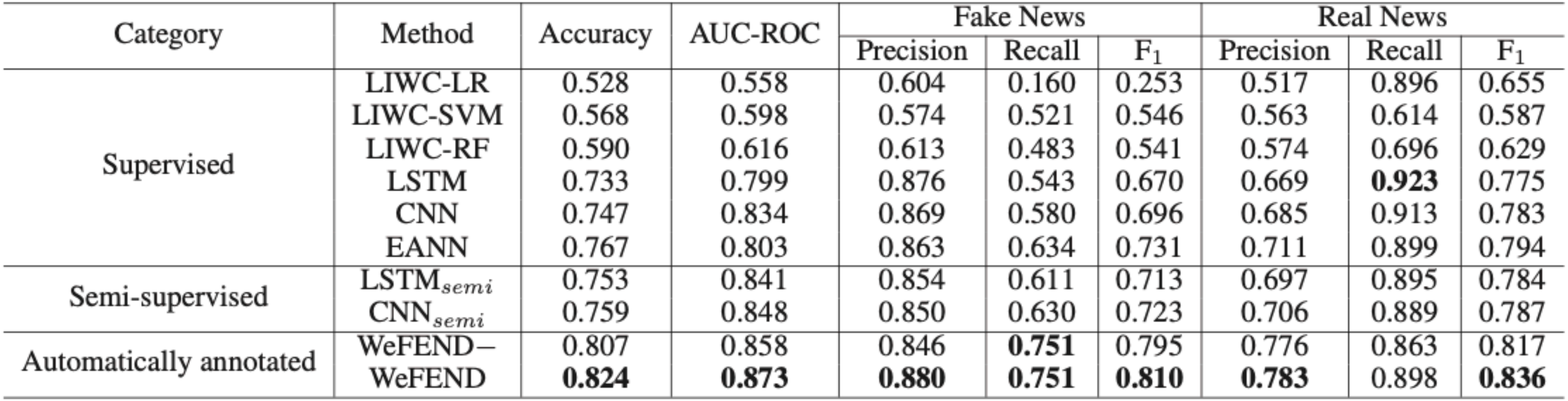


Category	Method	Accuracy	AUC-ROC	Fake News			Real News		
				Precision	Recall	F_1	Precision	Recall	F_1
Supervised	LIWC-LR	0.528	0.558	0.604	0.160	0.253	0.517	0.896	0.655
	LIWC-SVM	0.568	0.598	0.574	0.521	0.546	0.563	0.614	0.587
	LIWC-RF	0.590	0.616	0.613	0.483	0.541	0.574	0.696	0.629
	LSTM	0.733	0.799	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**

- LIWC-LR achieves the worst performance. The reason is that LIWC-LR is a linear model and hard to discriminate the complicated distributions of fake and real news content

Experiments

Performance Comparison

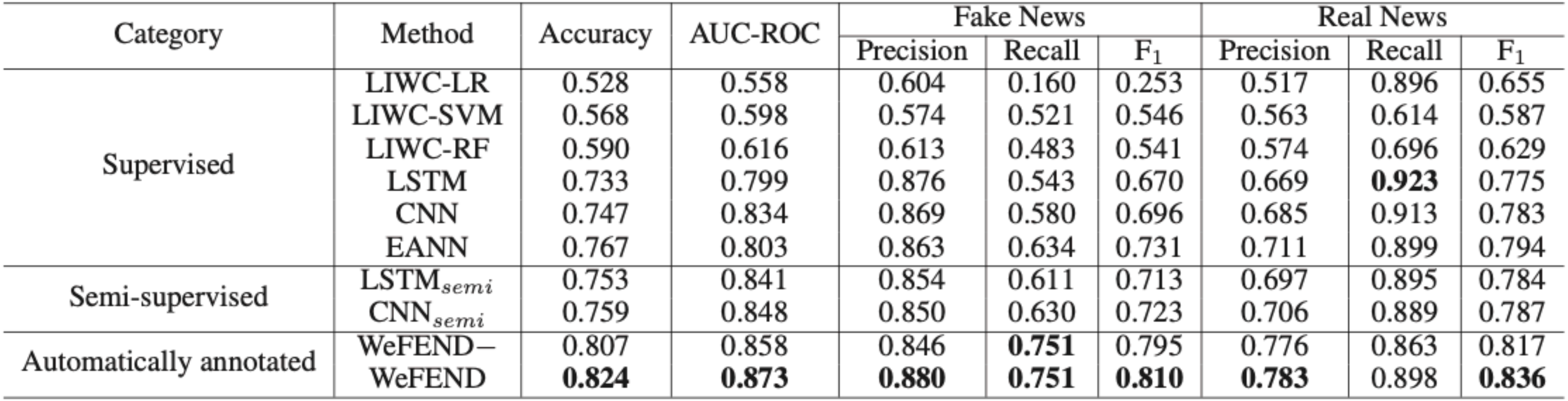


Category	Method	Accuracy	AUC-ROC	Fake News			Real News		
				Precision	Recall	F_1	Precision	Recall	F_1
Supervised	LIWC-LR	0.528	0.558	0.604	0.160	0.253	0.517	0.896	0.655
	LIWC-SVM	0.568	0.598	0.574	0.521	0.546	0.563	0.614	0.587
	LIWC-RF	0.590	0.616	0.613	0.483	0.541	0.574	0.696	0.629
	LSTM	0.733	0.799	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**

- In the semi-supervised setting, since the number of data largely increases (using unlabeled data to enlarges size of training set), the performance improvement in both models

Experiments

Performance Comparison

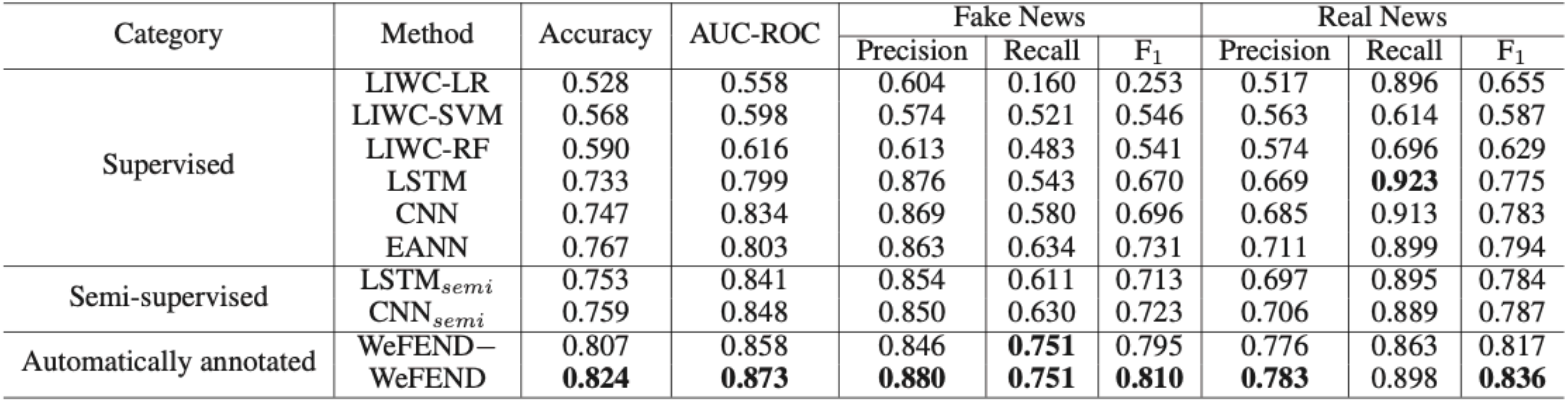


Category	Method	Accuracy	AUC-ROC	Fake News			Real News		
				Precision	Recall	F_1	Precision	Recall	F_1
Supervised	LIWC-LR	0.528	0.558	0.604	0.160	0.253	0.517	0.896	0.655
	LIWC-SVM	0.568	0.598	0.574	0.521	0.546	0.563	0.614	0.587
	LIWC-RF	0.590	0.616	0.613	0.483	0.541	0.574	0.696	0.629
	LSTM	0.733	0.799	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**

- The advantage of WeFEND is that it can automatically annotate unlabeled news, the performance of WeFEND– is better than models in the supervised and semi-supervised

Experiments

Performance Comparison



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**

- To reduce the influence of noisy labels, WeFEND has the data selector component based on reinforcement learning techniques, precision values of WeFEND are improved compared with WeFEND-

Experiments

Insight Analysis

- Aim to answer the following questions:
  - Does the distribution of news change with time?
  - Why should we use the reports to annotate the fake news?

Experiments

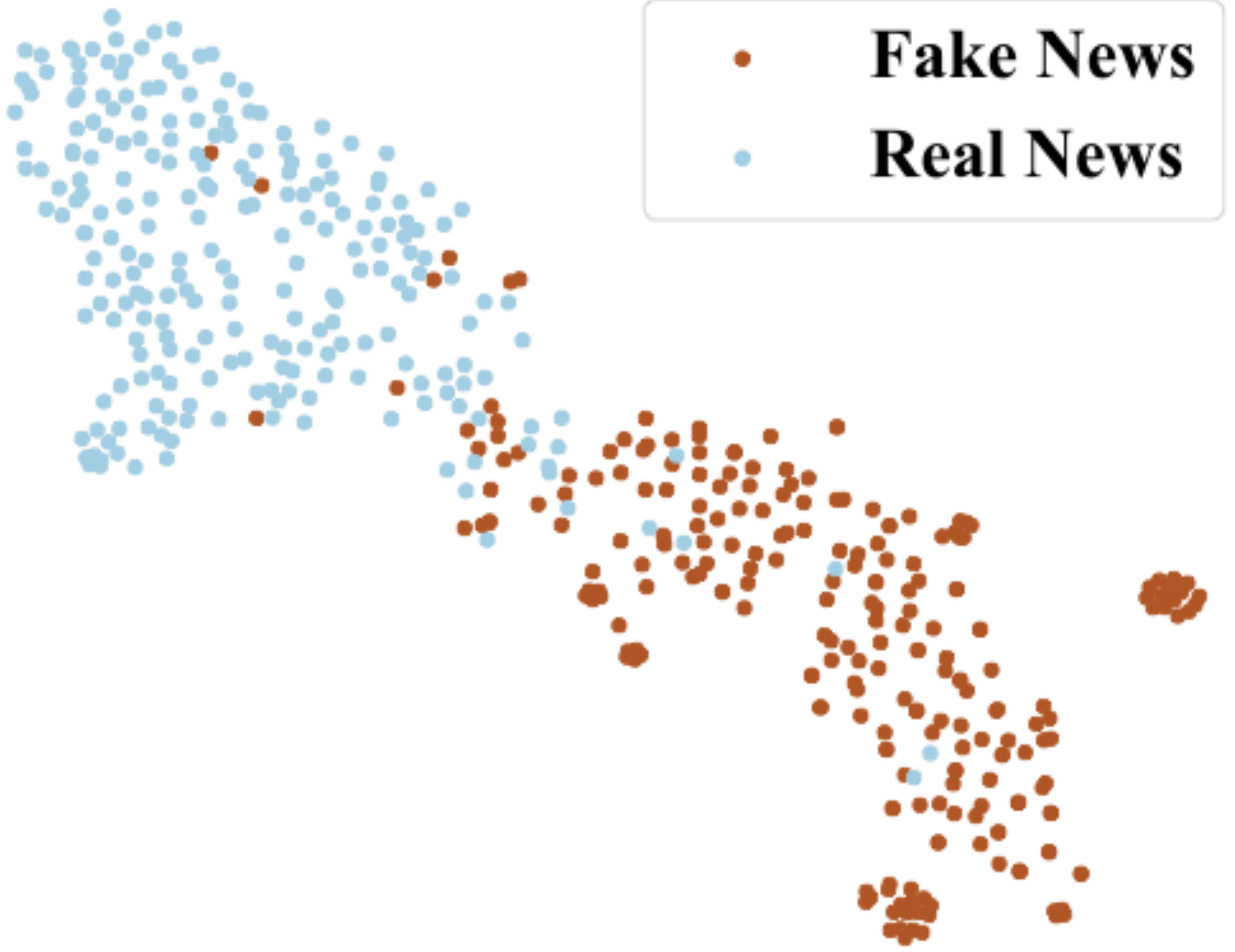
Insight Analysis

- Split the original training dataset consists of news content and reports into two sets:
    - 80% data as the new training set (denoted as D_t)
    - 20% data as the testing set for the same time window setting (denoted as D_s)
    - For the different time window setting, randomly select a subset samples from original testing dataset (denoted as D_d)
- The fake news detector and annotator are first trained on the news content of D_t, and then we separately test the models on D_s and D_d.

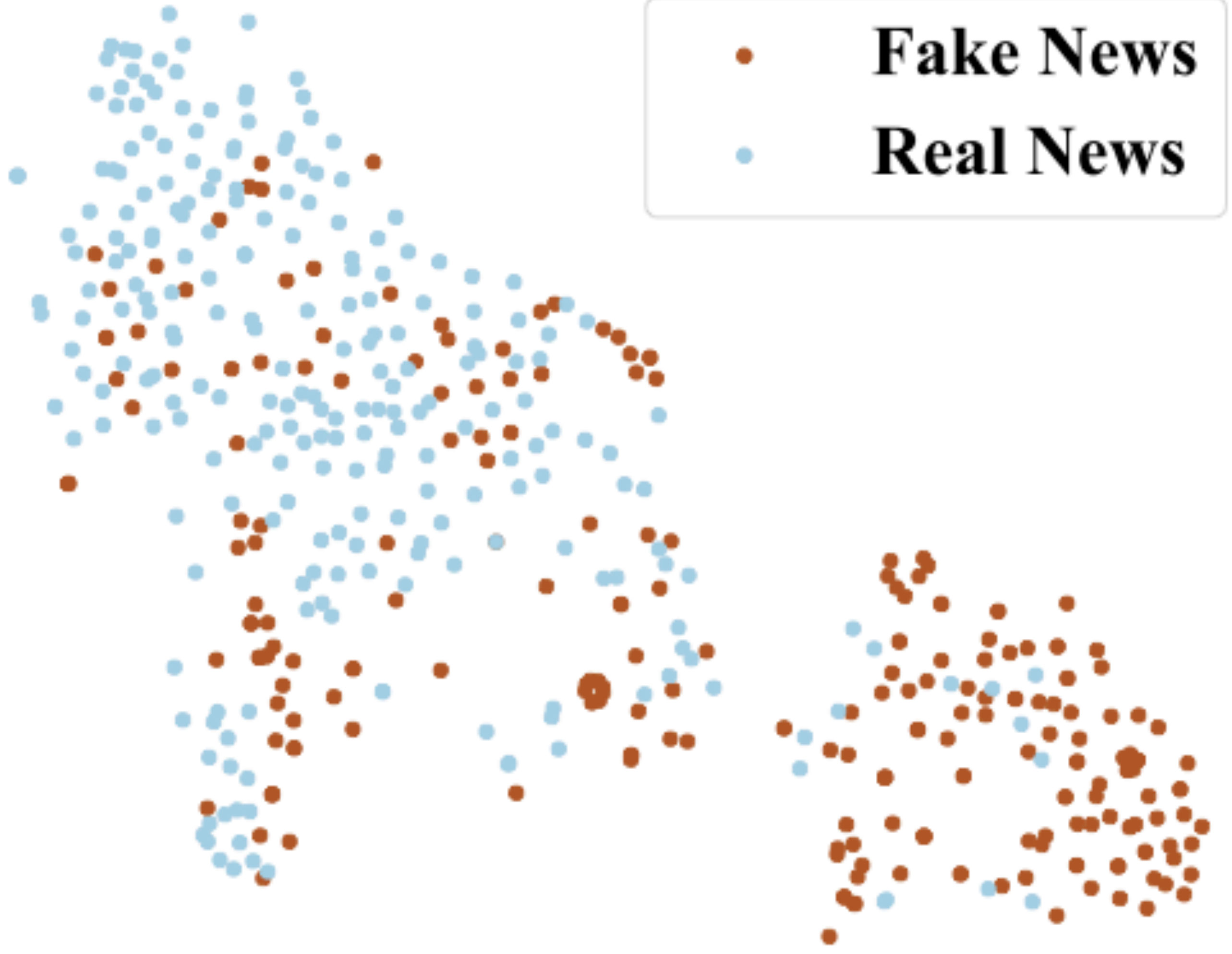
Experiments

Insight Analysis

- Visualizations of latent feature representations on D_s and D_d in t-SNE
  - D_s are very discriminative, and the segregated area between fake and real news is clear
  - D_d are twist together compared with D_s
- This comparison shows the feature representations of news in different time windows are significantly different with each other



(a) Same Time



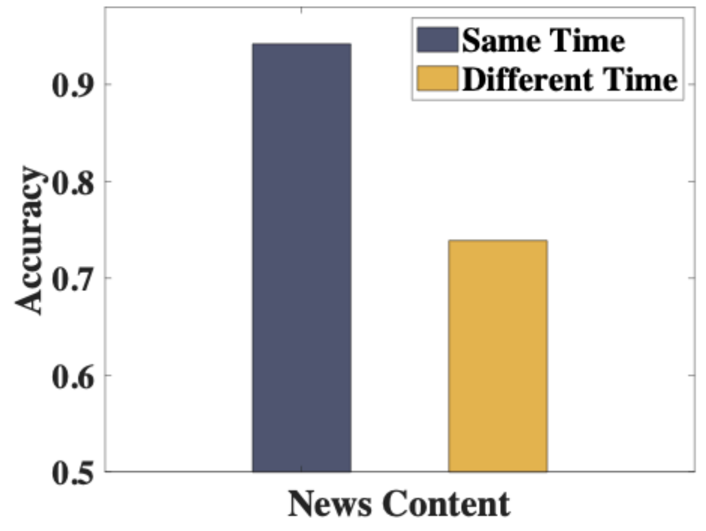
(b) Different Time

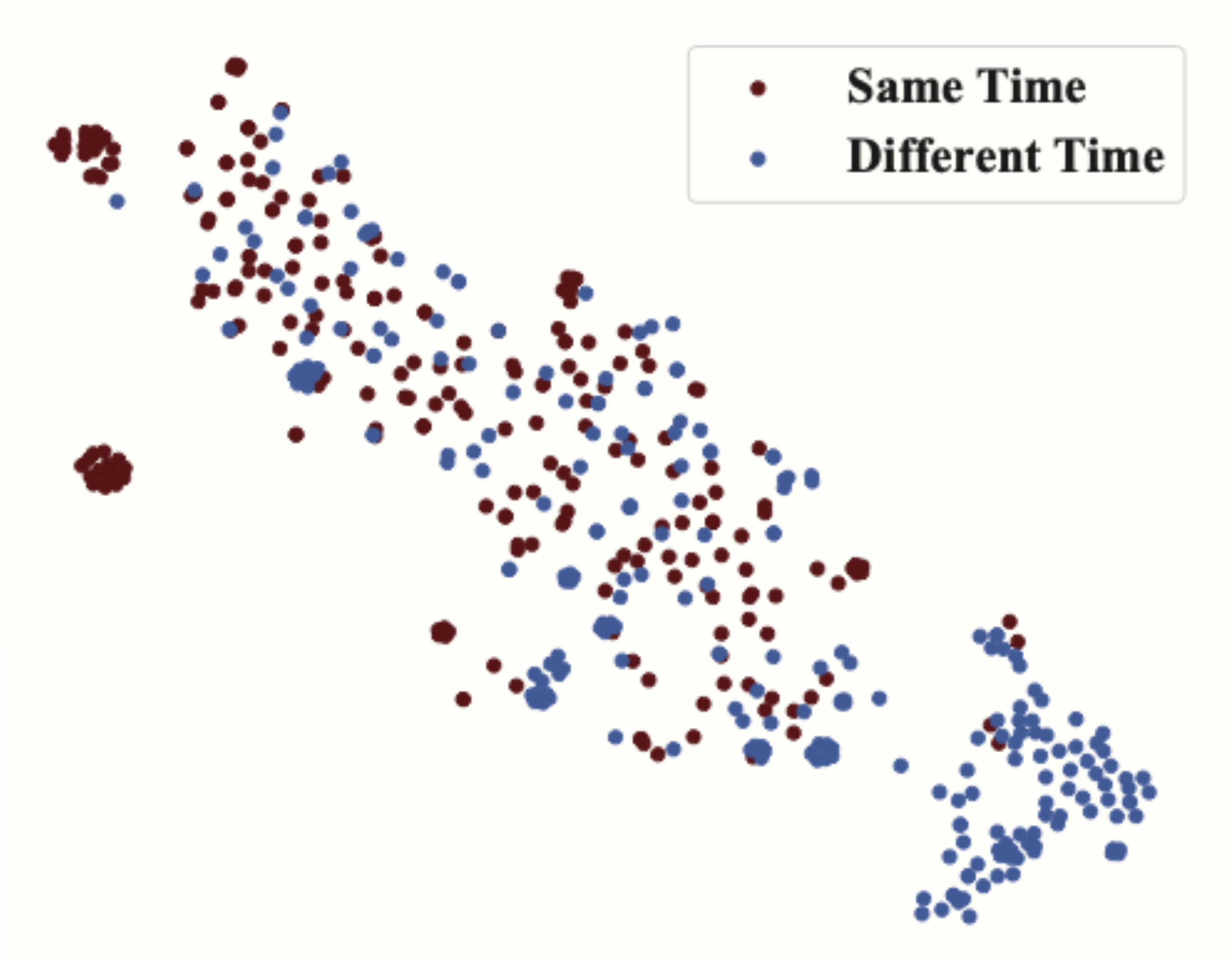
Experiments

Insight Analysis

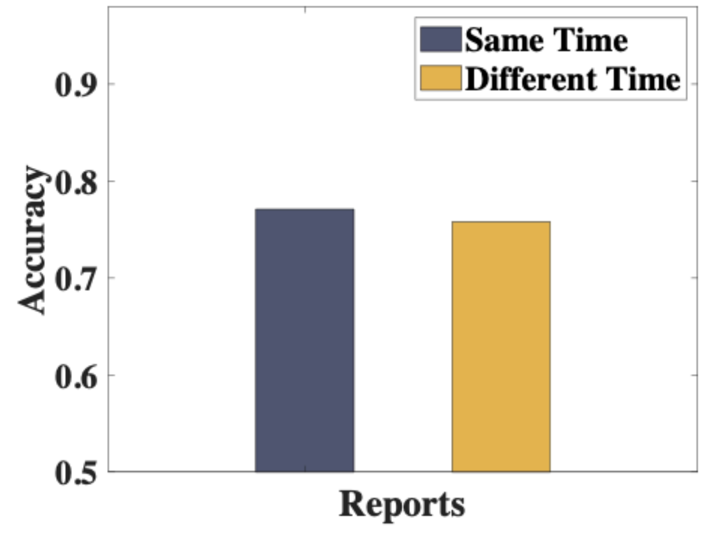
- The significant performance difference between two sets confirms that the distribution of news is changing

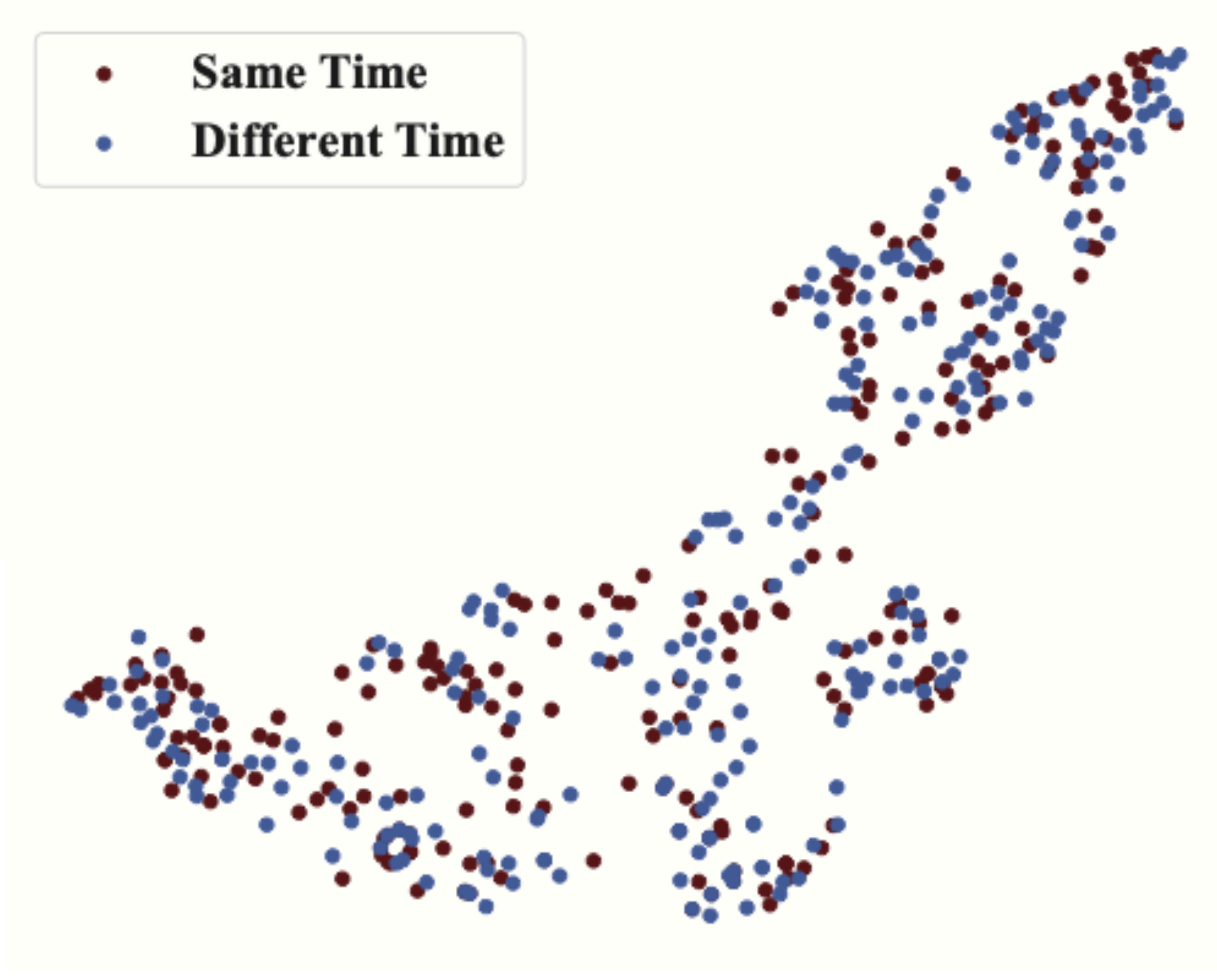
- The annotator can achieve similar performance on the same and different time set, demonstrates that the quality of fake news annotation on reports does not change with time





(a) News Content



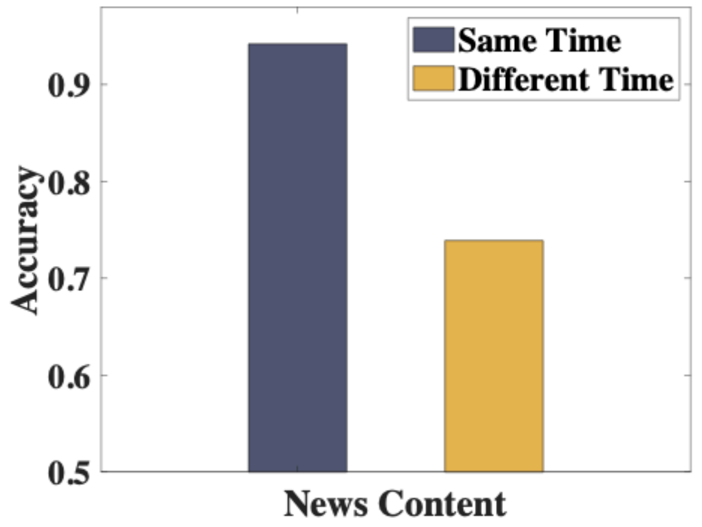


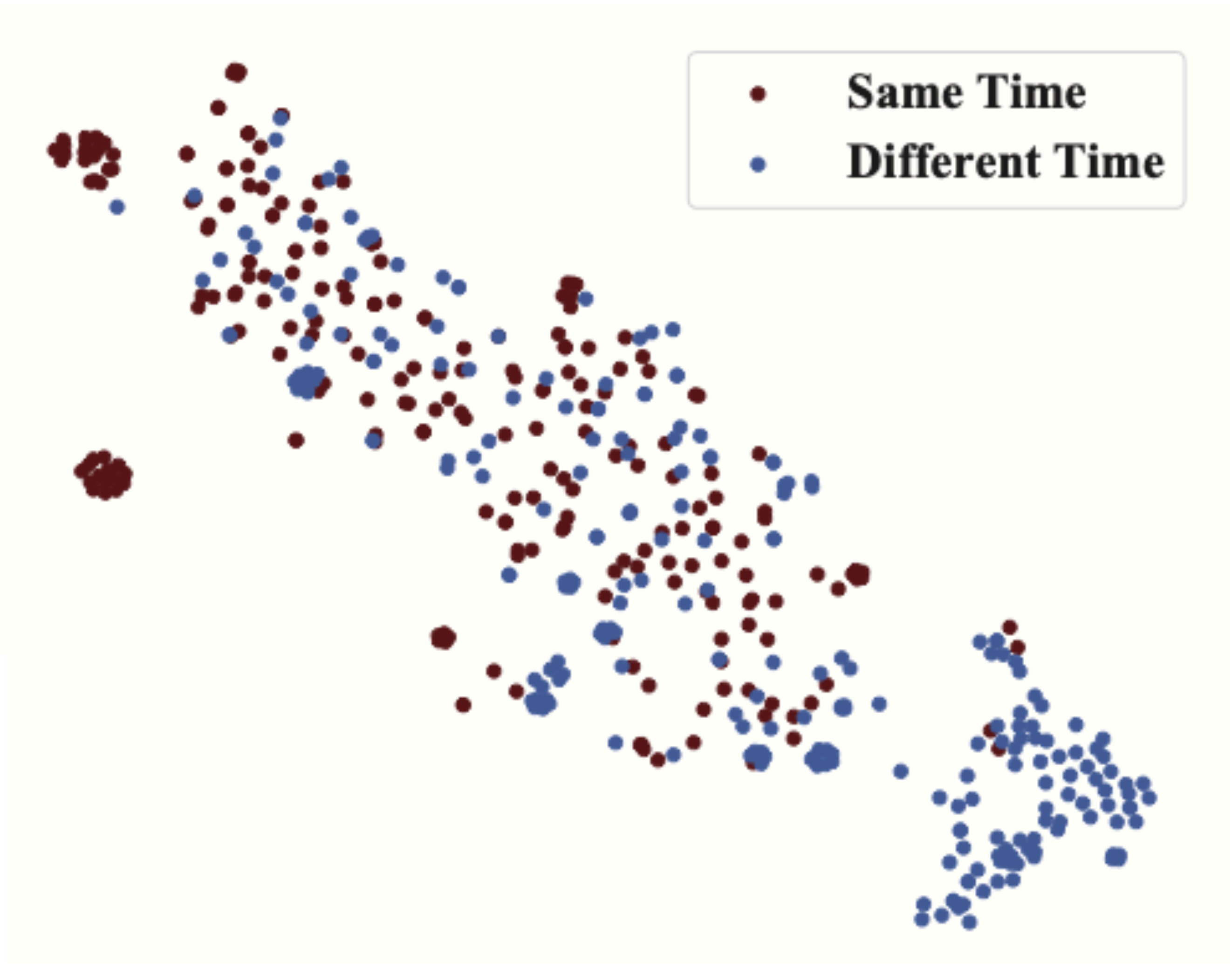
(b) Reports

Experiments

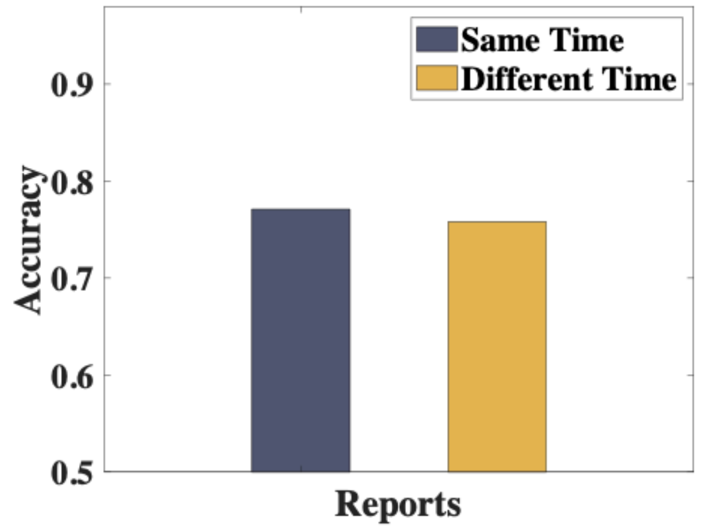
Insight Analysis

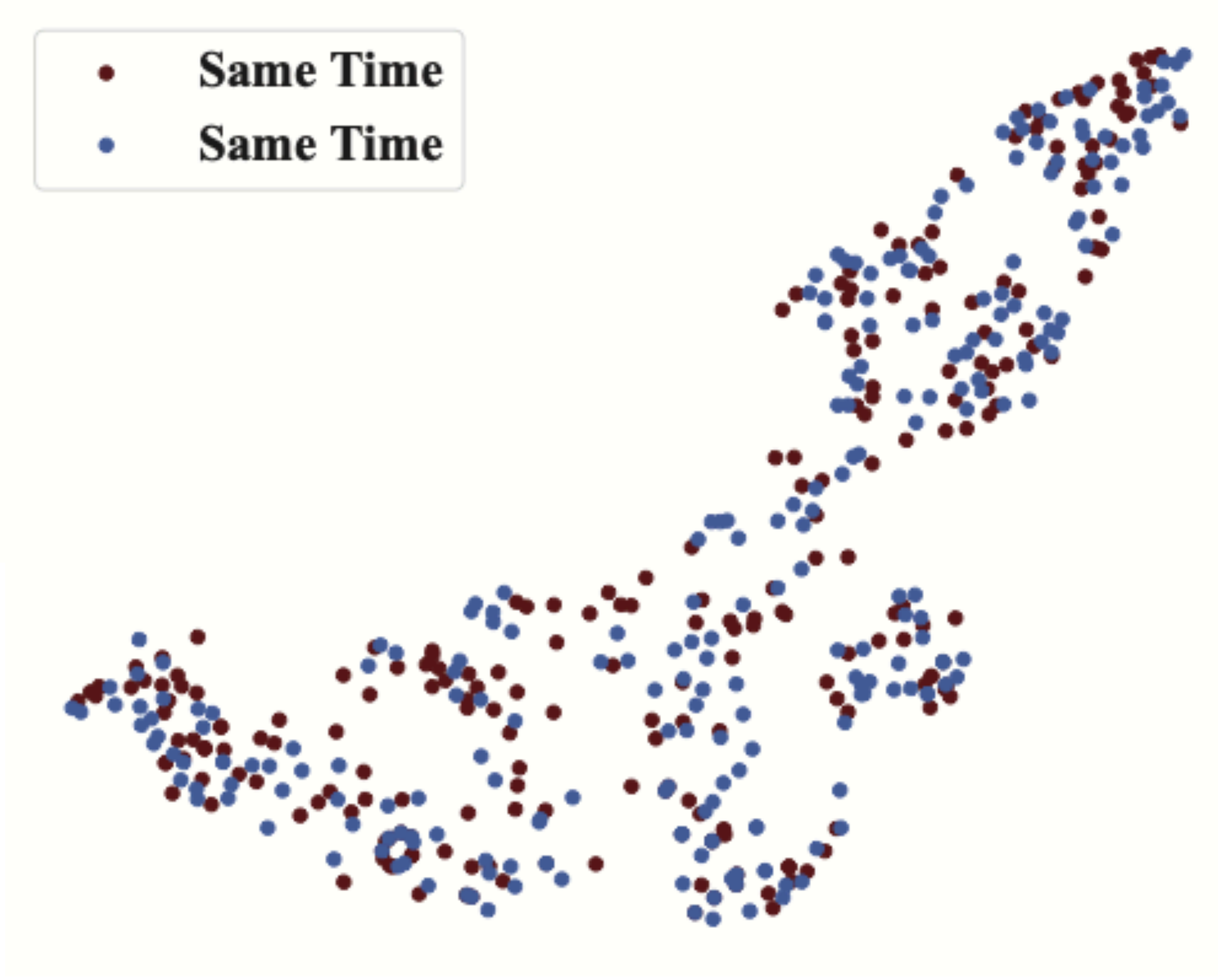
- Although the distributions of news content in the same time set and different time set have overlaps
  - The samples from two set are separately clustered at the top left and bottom right corner
  - This shows the distribution of news contents changes with time





(a) News Content



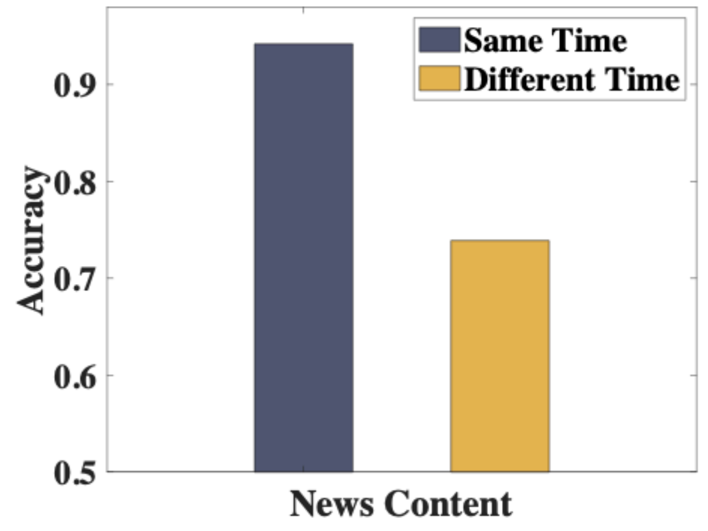


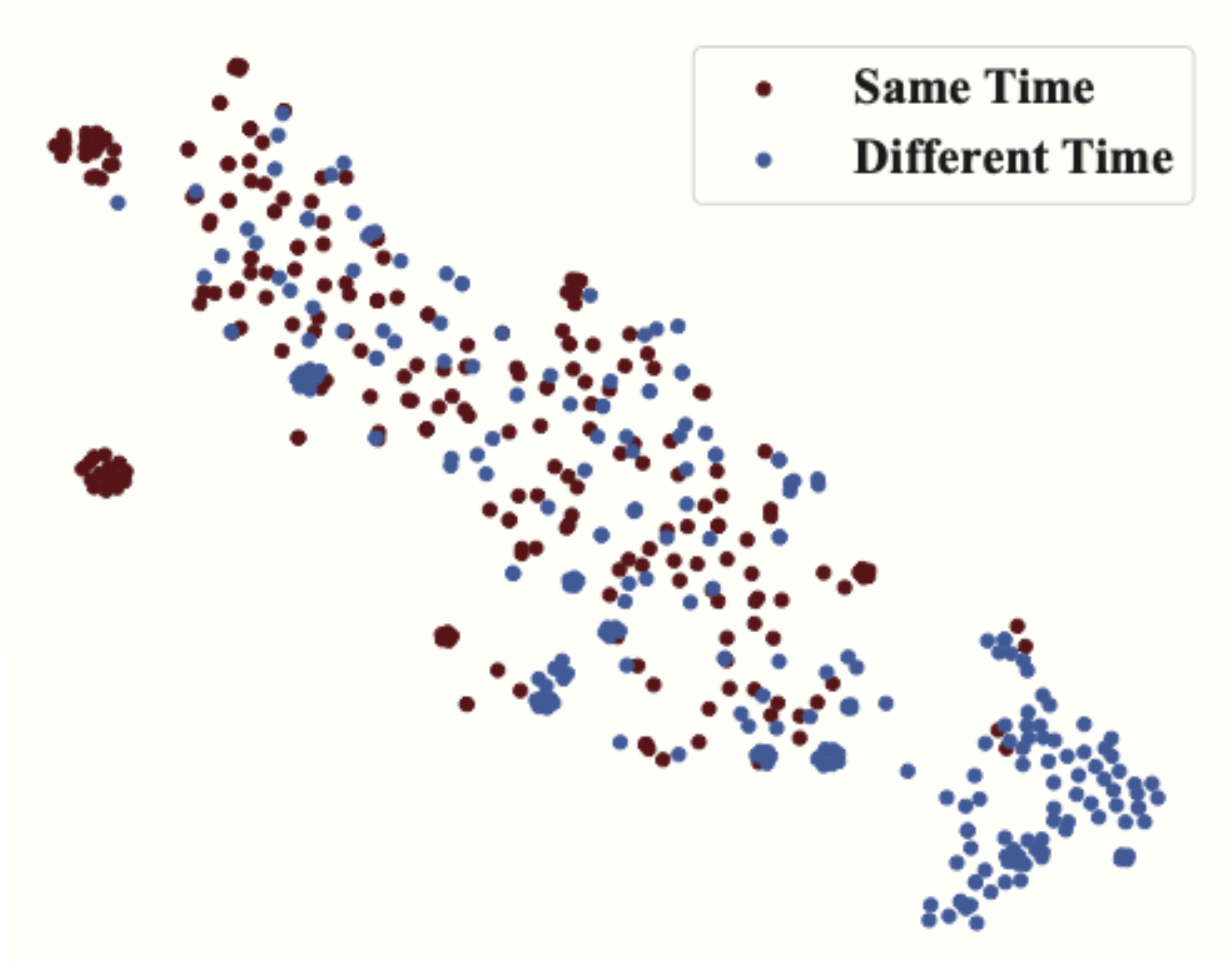
(b) Reports

Experiments

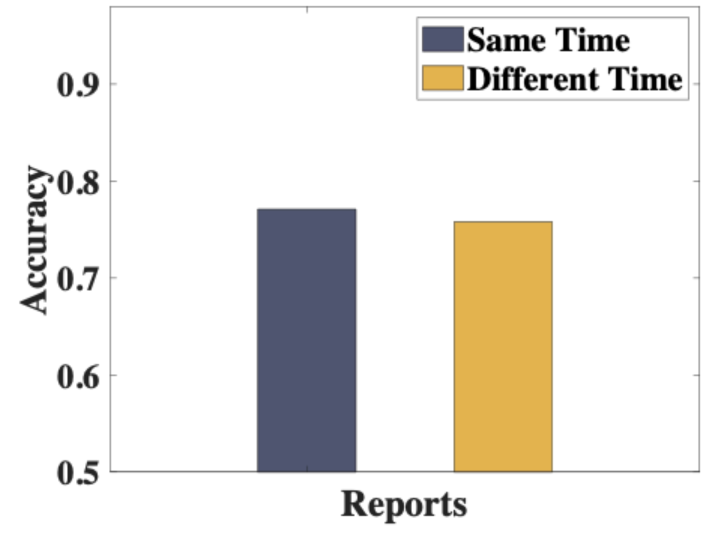
Insight Analysis

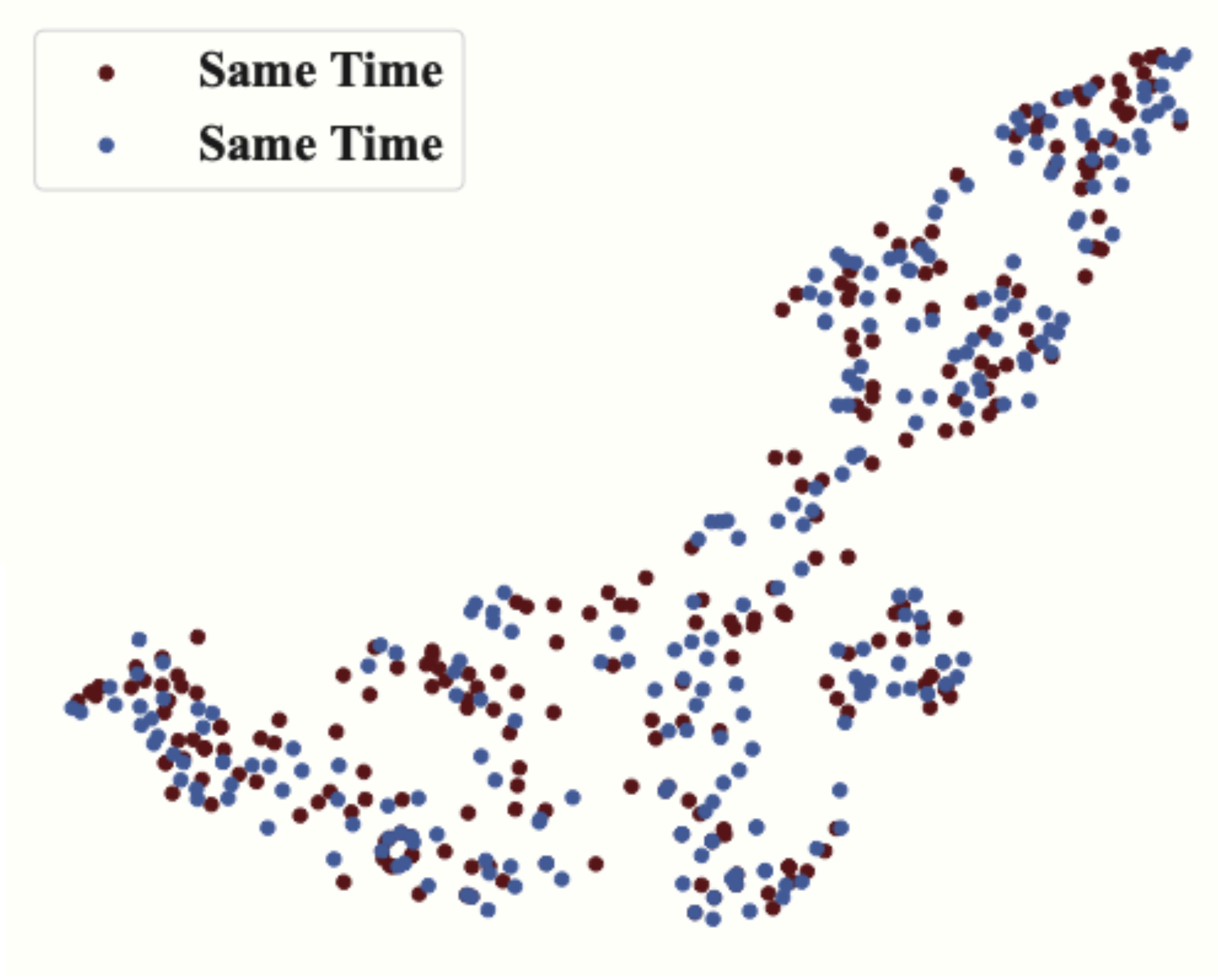
- In contrast, the representations of report messages from two sets are all twisted and cannot be distinguished
    - Proves that the distributions of reports is time invariant and further explains why the model trained on report messages achieves a consistent performance
    - Thus, the annotation based on reports can guarantee consistent quality even for fresh news articles





(a) News Content



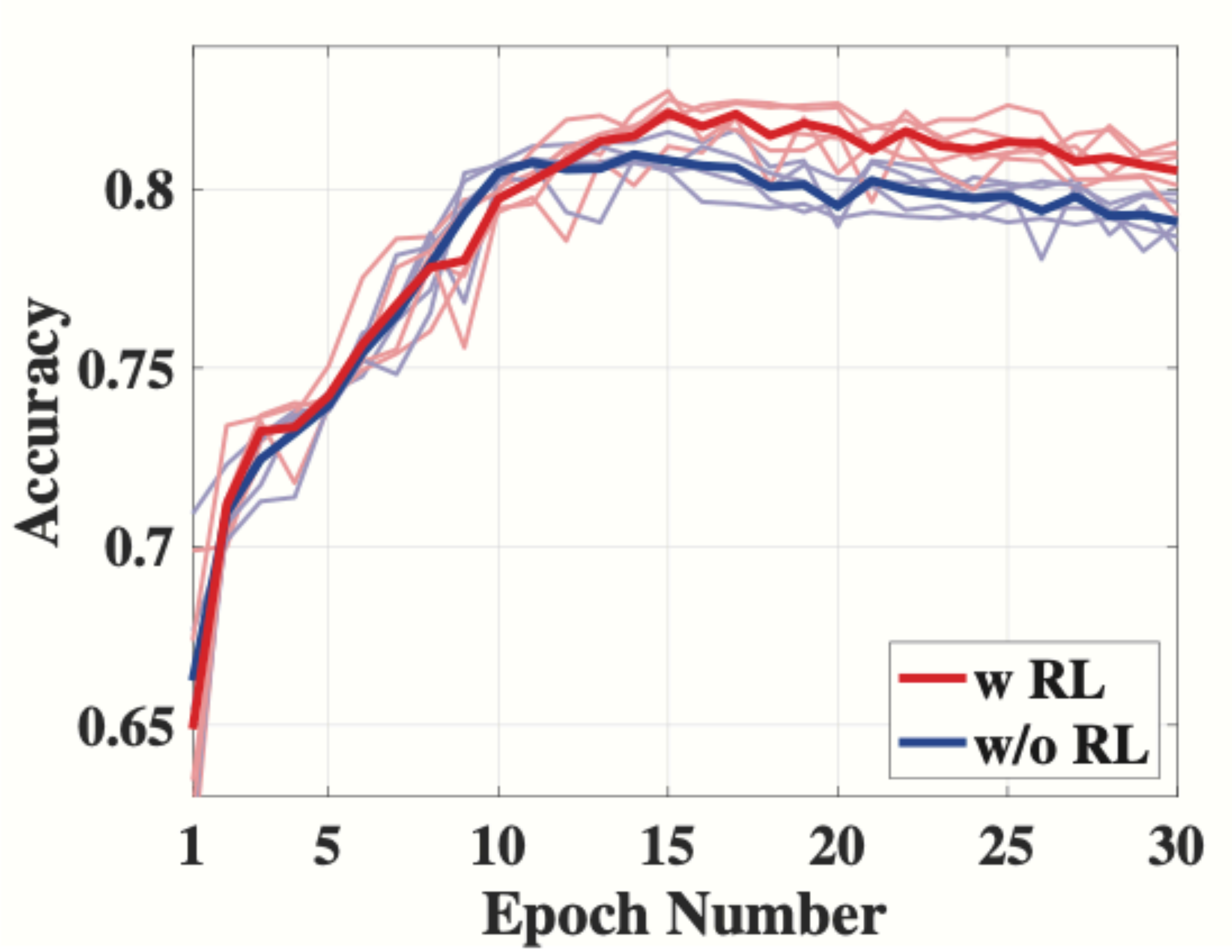


(b) Reports

Experiments

Insight Analysis

- As the probability output from fake news detection model can provide more information for the reinforced selector
  - Observe that the average accuracy of the model with reinforced selector is stably higher than that w/o reinforced selector after 12 epochs
- The ablation study shows that the designed reinforced selector is effective in improving the performance of fake news detection



Conclusion

- Proposed to investigate the important problem of fake news detection.
  - The dynamic nature of news make it infeasible to obtain continuously labeled high quality samples for training effective models
- Proposed a novel framework that can leverage user reports as weak supervision for fake news detection
- The reinforced selector based on reinforcement learning techniques chooses high-quality samples from those labeled by the annotator
  - By enhancing the quality and size of the training set, the proposed framework thus has shown significantly improved performance in fake news detection

Comments

of Reinforced Weakly-supervised FakE News Detection framework (WeFEND)

- Use time invariant of user reports to tackle the label newly emerging news problem
  - User report when occur user knowledge-domain problem ▲
- Use reinforcement learning method to filter high-quality training data
- Incorporate with image information may can improve the performance