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.