Uncertainty-Aware Reward-based Deep Reinforcement Learning for Intent Analysis of Social Media Information

Zhen Guo[1], Qi Zhang[1], Xinwei An[2], Qisheng Zhang[1], Audun Jøsang[3], Lance M. Kaplan[4], Feng Chen[5], Dong H. Jeong[6], Jin-Hee Cho[1]

[1]Department of Computer Science, Virginia Tech, VA, USA; [2]2810 Jackson Avenue, NY, USA;
[3]University of Oslo, Oslo, Norway; [4]DEVCOM Army Research Laboratory, MD, USA;
[5]University of Texas at Dallas, Richardson TX, USA; [6]University of the District of Columbia, DC, USA
zguo@vt.edu, qiz21@vt.edu, xan1@fordham.edu, qishengz19@vt.edu, audun.josang@mn.uio.no,
lance.m.kaplan.civ@army.mil, feng.chen@utdallas.edu, djeong@udc.edu, jicho@vt.edu

Abstract

Due to various and serious adverse impacts of spreading fake news, it is often known that only people with malicious intent would propagate fake news. However, it is not necessarily true based on social science studies. Distinguishing the types of fake news spreaders based on their intent is critical because it will effectively guide how to intervene to mitigate the spread of fake news with different approaches. To this end, we propose an intent classification framework that can best identify the correct intent of fake news. We will leverage deep reinforcement learning (DRL) that can optimize the structural representation of each tweet by removing noisy words from the input sequence when appending an actor to the long short-term memory (LSTM) intent classifier. Policy gradient DRL model (e.g., REINFORCE) can lead the actor to a higher delayed reward. We also devise a new uncertainty-aware immediate reward using a subjective opinion that can explicitly deal with multidimensional uncertainty for effective decision-making. Via 600K training episodes from a fake news tweets dataset with an annotated intent class, we evaluate the performance of uncertainty-aware reward in DRL. Evaluation results demonstrate that our proposed framework efficiently reduces the number of selected words to maintain a high 95% multi-class accuracy.

Introduction

Several recent studies (Apuke and Omar 2021; Shen et al. 2021) reported that fake news could be shared without bad intent, while it can contribute to a huge adverse impact on its propagation. Therefore, this work is motivated to provide a way of identifying the intent of fake news which can be used as the basis for effective intervention in mitigating fake news. More specifically, our intent classification method can contribute to identifying the right population and dealing with them differently to mitigate the impact of fake news propagation depending on their intent.

Text classification tasks have been studied by deep learning features of embedding and structure representation (Yogatama et al. 2017). Deep reinforcement learning (DRL) has been used to find an optimized structure representation by removing the noises, such as non-related words and lexical features (Zhang, Huang, and Zhao 2018). However, one of the main limitations in existing DRL-based text classification approaches is the use of a delayed reward to evaluate the prediction of the whole sequence of words. Because the final processing state can only be determined when all words are processed, this hinders the learning process as a DRL agent cannot receive an immediate reward upon its local decision.

This work tackles this issue by introducing a multidimensional uncertainty-aware reward in a DRL-based intent classifier. In addition, we propose an intent classification framework that can analyze the intents of fake and true news spreaders. In addition, we consider multidimensional uncertainty estimates to identify optimal parameters.

Our work makes the following **key contributions**:

1. We employ DRL algorithms to maximize the intent class prediction rate and minimize the number of words used in the intent classification process. Specifically, we introduce an immediate multidimensional uncertainty-aware reward to formulate an accumulated certainty reward.
2. We use the uncertainty estimates, such as vacuity and dissonance, in updating the current policy. This approach is the first to leverage a belief model, called *Subjective Logic*, that explicitly offers the capability of dealing with multidimensional uncertainty.
3. We resolve labor-intensive manual labeling tasks and annotate each news data either fake news or true news, by three annotators. Finally, we assign the dominant intent class to each news data piece to combine the labels collected from three sources.

Related Work

Intent Mining. A number of social science research (Apuke and Omar 2021; Koohikamali and Sidorova 2017; Shen et al. 2021) has studied users' motivations for sharing fake news and their intent. They found that the main reason for spreading fake news or false information was because they thought it was authentic and shared it unintentionally or even with good intent to help others or make fun. Recent studies have proposed several natural language processing (NLP) methods to do intent mining in different task domains, such as user intent mining from reviews (Khattak et al. 2021) and email intent (Shu et al. 2020).

DRL-based Text Mining. DRL has been explored in NLP classification tasks to select meaningful word tokens and maintain the sequential relationship of selected

words (Zhang, Huang, and Zhao 2018). That is, a better structure representation of whole text data can be learned by a DRL model to improve classification performance. The sentiment classification research (Jian et al. 2021; Wang et al. 2019) has used DRL to reduce noisy tokens from whole sentence to improve the accuracy of a prediction model. Unlike the works using sentiment labels, our work identifies the intents of news articles user intention identification from user reviews and emails. Hence, we manually label the intent classes from an existing fake news dataset. To the best of our knowledge, this work is the first that considers multi-dimensional uncertainty in the immediate reward of actions and quantifies the multidimensional uncertainty through a belief model called *Subjective Logic* (SL) (Jøsang 2016).

Intent Analysis Model

Intent Classes from Social Studies

In social sciences research, the intents of news spreaders, regardless of real or fake news, have been analyzed from the answers of questionnaires (Koohikamali and Sidorova 2017; Shen et al. 2021). However, no data-driven approaches have been used to study news spreaders' intents. In our manual annotations of news articles, we considers the following five intent classes based on social sciences findings:

- **Information sharing**: A common intent of online users' spreading behavior is purely sharing useful information to help other people. This intent includes sense-making or expertise sharing to facilitate the truth of shared information (Shen et al. 2021).
- **Political campaign**: This intent uses fake news to falsely perceive an opponent party's political figure or group to mislead public opinions to win elections (Purohit and Pandey 2019).
- **Socialization**: Online users share information to attract more friends and stay connected in online social networks (OSNs) (Apuke and Omar 2021). They can expand his/her social cycle by sharing news, often leading to finding common and interesting topics.
- **Rumor propagation**: This intent misleads users by sharing a rumor or unverified information (Purohit and Pandey 2019). Rumors can alter users' emotions and attitudes toward certain events and increase uncertainty.
- **Emotion venting**: Online users can propagate fake news when they feel emotional happiness or disturbances (e.g., anger, depression, sadness) from reading good or bad events (Alsmadi, Alazzam, and AlRamahi 2021).

Long Short-Term Memory (LSTM) Intent Classifier

Since the current dataset includes a collection of tweets, we call one piece of news data "a news tweet" for consistency. First, we manually label each fake or true news tweet with an intent class in the Experiment Setup section. Accordingly, this label is a gold intent class $\mathbf{y}$ for a news tweet $\mathbf{x}$. Then, a classifier is learned from the annotated dataset to predict one's intent.

Model Components. This language model has a recurrent structure of k embedding layers (θ_{em}) and k LSTM cells (θ_{lstm}) to process a sequence of k input words in a news tweet $\mathbf{x} = \{x_1, \ldots, x_k\}$. After k iterations with the final state h_k, a critic network (θ_{critic}) with dense linear layers and ReLU and softmax activation functions can predict a distribution of P intent classes. We borrow the term 'critic' to highlight the role in the DRL model. As in a traditional LSTM cell at iteration t, the inputs are a cell vector c_{t-1}, a hidden state vector h_{t-1}, and the input word vector w_t embedded from a word token x_t. The outputs are new c_t and h_t sent to the next iteration LSTM cell.

Loss Function. A parameter set θ_{IC} of the three components in the intent classifier is updated from the training step, which minimizes the *cross-entropy loss* with the known gold intent labels. Considering the over-fitting prevention, the L2 regularization term $||\theta_{IC}||_2^2$ is added to this loss by:

$$\mathcal{L}_{IC} = -\sum_{y=1}^{P} \hat{p}(y, \mathbf{x}) \log p(y|\mathbf{x}, \theta_{IC}) + \alpha||\theta_{IC}||_2^2, \quad (1)$$

where α is the regularization rate, $p(y|\mathbf{x}, \theta_{IC})$ is predicted from the final softmax layer, and $\hat{p}(y, \mathbf{x})$ is the one-hot distribution of the gold label $\mathbf{y}$, which has P values with a single high as 1 and all others are 0.

DRL-based Intent Classifier

This section discusses the sentence structure optimization steps by our policy gradient-based DRL (see Figure 1 for detail). Keeping the temporal relationship of the intent-related words can improve the intent prediction from the LSTM intent classifier described in the previous section. Since the previous LSTM classifier serves as a fixed environment to support the state transitions, their parameters θ_{IC} are all frozen during the DRL learning of θ.

Model Steps. The DRL model adds a word selection step to the LSTM classifier to form a time series of actions $\tau = \{a_1, \ldots, a_k\}$ from each 'keeping' or 'masking' decision of the actor network. Then, the selected or non-masked words generate an optimized input sequence $\mathbf{x}' = \{x_1, x_2, \ldots, x_{k'}\}$ of k' words ($k' \leq k$). Finally, this shorter $\mathbf{x}'$ is processed by the three components of the LSTM intent classifier to predict the gold intent with higher accuracy. A training episode in DRL stands for one trial of the Markov decision process (MDP) by masking non-intent words from steps 1 to k of a news tweet. In addition, an episode is applied based on only one news tweet data $\mathbf{x}$. It is obvious that the golden class $\mathbf{y}$ for tweet $\mathbf{x}$ can be achieved during our annotation step. The truth intent class $\mathbf{y}$ (e.g., socialization) can provide a delayed reward to lead the DRL to a higher reward by repeating five episodes in one training epoch.

Word Selection DRL. By setting $\mathbf{x}' = \mathbf{x}$, the DRL model extends an actor network to each LSTM cell θ_{lstm} in the previous LSTM intent classifier with the following details:

- **State**: At each step t, an input feature vector w_t from the embedding layer θ_{em} represents the tokens of an input word x_t. The LSTM cell θ_{lstm} processes w_t with c_{t-1}

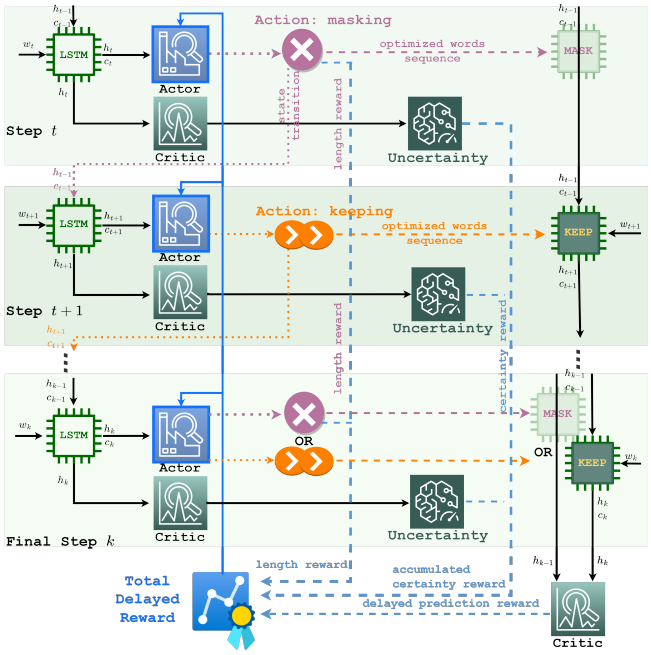

Figure 1: The overview of the DRL model with an uncertainty-based immediate reward.

and h_{t-1} and passes c_t and h_t. This hidden state vector h_t serves as the state s_t for the actor network at the current step as $s_t = \pi(w_t, c_{t-1}, h_{t-1}; \theta_{lstm})$.

- **Actor**: The actor network has one dense linear layer followed by a ReLU activation and another dense linear layer followed by softmax activation of two nodes. The parameters of all the neurons are in θ and can be trained by the DRL. Given a state s_t, the actor generates a policy $\pi(a_t|s_t, \theta)$ for two actions.
- **Action**: If an actor chooses $a_t = 1$, the 'keeping' action maintains x_t in the optimized input sequence $\mathbf{x}'$. Otherwise, if the actor chooses $a_t = 0$, the 'masking' action removes x_t from $\mathbf{x}'$.
- **State Transition**: The state transition determines how each action a_t controls the state s_{t+1} for the next step. Since s_{t+1} is generated from the next LSTM cell, a_t decides the inputs for the LSTM cell of step $t+1$ by:

$$s_{t+1} = \begin{cases} \pi(w_{t+1}, c_{t-1}, h_{t-1}; \theta_{lstm}) & \text{if } a_t = 0 \text{ 'masking'}; \\ \pi(w_{t+1}, c_t, h_t; \theta_{lstm}) & \text{if } a_t = 1 \text{ 'keeping'}, \end{cases} \quad (2)$$

Delayed Reward. A *delayed reward* is from the intent class distribution of the whole optimized input sequence $\mathbf{x}'$ of one news tweet, which is calculated from the LSTM classifier after the last word $w_{k'}$. This delayed reward is a unique feature in some NLP problems when common immediate reward is unavailable. Considering the prediction of the true intent class label, as $R_{pred} = p_{\theta_{critic}}(\mathbf{y}|\mathbf{x}')$, and the masked length reward of input words, the delayed reward is defined by:

$$R = p_{\theta_{critic}}(\mathbf{y}|\mathbf{x}') + \lambda(k - k')/k, \quad (3)$$

where θ_{critic} is a parameter for the critic network in the LSTM intent classifier, $\mathbf{y}$ is the golden label for a news tweet $\mathbf{x}$, k' is the length of selected words in $\mathbf{x}'$, and λ is a weight.

Policy Gradient. The goal of our DRL model is to maximize the delayed reward in Eq. (3). Similar to many other NLP problems where an immediate reward is unavailable after deciding an individual step's action, we will learn the actor's policy by policy gradient (PG)-based models. For example, the current stochastic policy maximizes the expected delayed reward by calculating the gradients $R \log \pi(a_t|s_t, \theta)$ in REINFORCE algorithm (Williams 1992). Based on the gradients for all k transition steps in one episode following PG's on-policy property, REINFORCE calculates the *negative log loss* of the actor θ's policy as:

$$\mathcal{L} = -\sum_t R \log \pi(a_t|s_t, \theta), \quad (4)$$

Uncertainty-based Immediate Reward

In each local step t of the DRL model, when the hidden vector $h_t = s_t$ is passed directly to the critic network θ_{critic}, the intent class distribution $p(s_t, \theta_{critic})$ from the critic network can be represented by Subjective Logic (Jøsang 2016). Owing to the uncertainty metrics in the SL-based opinion, we can access the accumulated uncertainties or certainties from each k decision steps in one training episode.

Formulating SL-based opinion from local intent classification probabilities. Remind that we adopt SL to consider multidimensional uncertainty where a traditional Dirichlet probability density function (PDF) can be easily mapped to a multinomial opinion which can provide a way to estimate different types of uncertainties. These local intent probabilities $p(s_t, \theta_{critic})$ can be regarded as a multinomial opinion in SL corresponding to five beliefs towards the intent classes. The SL opinion considers uncertainty metrics, such as *vacuity* caused by a lack of evidence and *dissonance* introduced by conflicting evidence. Since the local intent probabilities cannot reflect the level of vacuity, we apply the vacuity maximization technique (Jøsang 2016) on the local probabilities of P classes and generate a vacuity maximized opinion $\omega_t = [\mathbf{b_t}, \mathbf{u_t}, \mathbf{g}]$. There are P belief masses in SL opinion as $\mathbf{b_t}$. The $\mathbf{u_t} = [u_t^{vac}, u_t^{dis}]$ refers to two considered uncertainty metrics. The base rates in vector $\mathbf{g}$ are the distribution of P classes in annotated set $\mathcal{D}$.

Total Delayed Reward based on Accumulated Certainty Estimates. Based on the previous discussion of the critic's immediate certainty metrics collected from local steps, the accumulated certainty metric serves as a new component of the total delayed reward by extending Eq. (3) as:

$$R_{delay} = p_{\theta_{critic}}(\mathbf{y}|\mathbf{x}') + \lambda(k - k')/k + \beta \sum_t (1 - u_t). \quad (5)$$

where u_t is either vacuity u_t^{vac} or dissonance u_t^{dis}, and β is the weight of accumulated certainty reward.

Experiment Setup

Datasets. We use the publicly available dataset *LIAR 2015* (Wang 2017), which has 2,511 fake and 2,073 real news tweets verified by fact-checking agencies.

Table 1: MODEL HYPER-PARAMETERS SETTING

Model	Parameter	Value	Parameter	Value
LSTM	Training epochs	15	Padding words length k	20
	Learning rate	0.0003	Dimensions of h_t and c_t	128, 128
	Dropout rate	0.25	Dense layers dimensions	[257, 5]
	Batch size	32	Embedding dimensions	128
DRL	Training epochs	100	Dense layers dimensions	[257, 2]
	Learning rate	0.01	Mini-batch episodes	5
	Batch size	32	Length reward weight λ	0.5

Intent annotation. Each news tweet is manually annotated by three annotators to give one of the five intent classes discussed. Then, 33% of all news is annotated as a dataset $\mathcal{D}$ of 1,500 tweets, with 835 fake and 665 true news. Finally, we assign the dominant intent class from the three sources to each news tweet. The base rate $\mathbf{g}$ for five classes are: 'information sharing': 0.423, 'political campaign': 0.275, 'socialization': 0.135, 'rumor propagation': 0.086, and 'emotion venting:' 0.081.

Data preparation. The annotated dataset $\mathcal{D}$ is split into fake news only, true news only, and both. By training LSTM and DRL models, we use 80% as a training set and 20% as a testing set. We use prefix padding of length $k = 20$.

Parameterization. Our proposed DRL model aim to improve intent prediction by optimizing sentence representation structures, which will cause a reduction of words from the input data of the LSTM intent classifier. In addition, we discuss the role of uncertainty-aware rewards by the SL in intent prediction. Table 1 summarizes the model hyper-parameters and their default settings.

- LSTM non-RL model (named 'LSTM') learns the parameters θ_C based on the loss function in Eq. (1).
- REINFORCE (named 'DRL', from section DRL-based Intent Classifier section) is a basic PG method based on the delayed reward Eq. (3) and loss function Eq. (4).
- REINFORCE with uncertainty-aware reward (named 'DRL-CV' and 'DRL-CD' for vacuity and dissonance) uses the loss function in Eq. (4) by replacing R with Eq. (5).

Metrics. Our DRL framework is evaluated by measuring (1) the multi-class classification accuracy counted by the ratio of correctly predicted data over the total of the testing data; (2) the effectiveness of DRL by prediction of the gold intent class from the critic network; and (3) the efficiency for DRL represented by the length of the optimized sequence.

Experiment Results & Analysis

During one DRL training epoch of each news tweet, we collect the states, actions, and rewards by running five mini-batch episodes. Thus, each epoch contains $6,000$ training episodes. Our models can converge within 100 epochs and there are a maximum of 600K episodes. In the testing of DRL models, the actors always follow the policy strictly.

Multi-class Classification Accuracy

Although this accuracy is not directly optimized in the DRL algorithm and rewards, it is crucial for our intent classification goal. We list the accuracy scores for five classes and each intent class, along with the gold class prediction R_{pred} in Table 2. The highest accuracy scores for DRL models with different weights λ are within the same range, so we only show the case of $\lambda = 0.5$ to compare to the DRL with certainty reward models. The individual intent classes in the last column follow the same order in $\mathbf{g}$. The basic DRL model improved the gold class prediction by 5.3% and increased the intent class classification by 11% for all news data. The LSTM model has similar accuracy for fake and true news, but all DRL models classified fake news with higher accuracies than true news. Although the overall accuracy under our DRL-CV and DRL-CD is similar, the new certainty reward shows a slight improvement for the annotated intent class prediction. Also, in general comparison, the two uncertainty-based DRL models can increase the accuracy for classes 1 'information sharing' and 2 'political campaign' but decrease the prediction of classes 3 'socialization' and 4 'rumor propagation'.

Table 2: MULTI-CLASS ACCURACY FROM TESTING

Model (λ, β)	Test Data	Gold Class Prediction	Acc.	Acc. by Intent Class
LSTM	Fake	0.806	0.866	[0.853, 0.852, 0.727, 0.941, 1.0]
	True	0.840	0.867	[0.898, 0.727, 0.889, 0.714, 1.0]
	Total	0.820	0.860	[0.871, 0.802, 0.792, 0.850, 1.0]
DRL (0.5, N/A)	Fake	0.863	0.978	[0.941, 1.0, 1.0, 1.0, 1.0]
	True	0.887	0.958	[0.966, 0.909, 1.0, 0.857, 1.0]
	Total	0.873	0.970	[0.951, 0.963, 1.0, 0.943, 1.0]
DRL-CV ($0.5, 0.01$)	Fake	0.871	0.983	[0.956, 1.0, 1.0, 1.0, 1.0]
	True	0.886	0.950	[0.966, 0.955, 0.944, 0.714, 1.0]
	Total	0.877	0.970	[0.960, 0.982, 0.978, 0.886, 1.0]
DRL-CD ($0.5, 0.05$)	Fake	0.874	0.983	[0.956, 1.0, 1.0, 1.0, 1.0]
	True	0.890	0.958	[0.966, 0.955, 1.0, 0.714, 1.0]
	Total	0.880	0.973	[0.960, 0.982, 1.0, 0.886, 1.0]

Basically, DRL models show a high multi-class accuracy of 97% across all settings. Our DRL models aim to maximize the prediction accuracy while reducing the length of the optimized sequences. Hence, in the next sections, we use at least 95% overall multi-class accuracy performance to check the achieved minimum length for each DRL model.

Effectiveness Reward by Prediction

Our reward function in Eq. (5) can maximize the effectiveness metric R_{pred} for the direct gold class prediction. We show this metric from the training episodes and the testing data, and compare it with the total delayed reward in Figure 2. For the basic DRL models, when the weight λ of the removed length is higher, both the training and testing effectivenesses decrease while the total reward still increases. This means the loss of effectiveness is compensated by the length reward. However, the uncertainty-related reward can increase training and testing effectiveness, compared to the DRL model with the same weight $\lambda = 0.5$. In addition, the effectiveness of true news data is improved with the help of the certainty reward.

Efficiency by Optimized Length

The masked length reward in Eq. (5) is DRL's main contribution to reducing the number of noisy words in the news

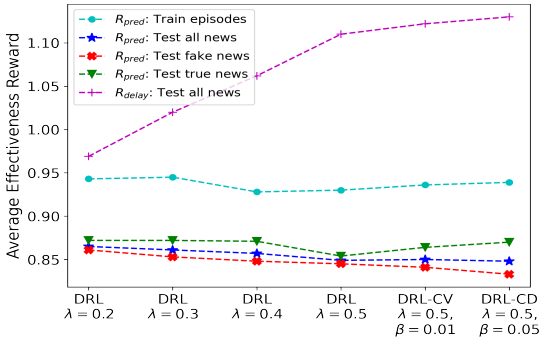

Figure 2: Effectiveness scores from DRL models.

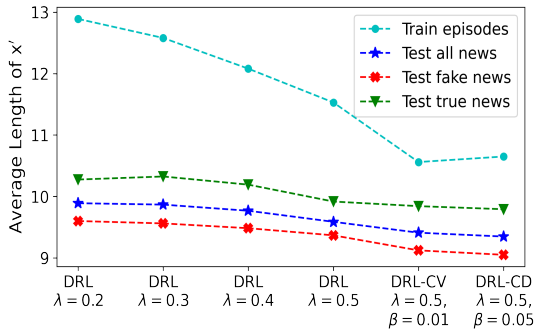

Figure 3: Efficiency metric from DRL models.

tweets. Figure 3 illustrates the minimum number of words to maintain an overall classification accuracy of 95%. Both training and testing lengths are reduced by adding a larger weight λ for the masked length, meaning that the DRL models maintain 95% classification accuracy while keeping less relevant words. Fake news in testing removes more words than true news, leading to a higher reward. The two uncertainty-related models reduce the length of both training and testing, indicating that the additional small amount of certainty values, both vacuity and dissonance, can help the DRL agent to explore more to the status of a shorter length. The decrease of the kept length of DRL-CV and DRL-CD, compared to DRL with $\lambda = 0.5$, can increase the efficiency reward and finally achieves a higher total delayed reward, as shown by the magenta line in Figure 2.

Conclusions

We annotated the intent class for the fake and true news from the dataset LIAR15. We used this annotated dataset to train the LSTM intent classifier, and then created a DRL model to help reduce the noisy words. The DRL model with optimized structure representation greatly improved the multi-class classification accuracy from the pretrained LSTM intent classifier. We then added an uncertainty-aware reward to help the DRL model reduce the length of words further, while maintaining the high level of intent class classification accuracy. Our findings prove that two uncertainty metrics, vacuity and dissonance, can help the DRL agent's training to reach shorter lengths and higher the gold class predictions. The two uncertainty-aware reward models can also decrease the length of testing data and increase the effectiveness of true news testing data, resulting in attaining a higher total delayed reward for the total test data. We will dig into more relevant metrics for the rewards by effectiveness and efficiency in our future work.

Acknowledgments

This work is partly supported by the Army Research Office under Grant Contract Number W91NF-20-2-0140 and NSF under Grant Numbers 2107449, 2107450, and 2107451.

References

Alsmadi, I.; Alazzam, I.; and AlRamahi, M. A. 2021. An ontological analysis of misinformation in online social networks. *arXiv:2102.11362*.

Apuke, O. D.; and Omar, B. 2021. Fake news and COVID-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, 56: 101475.

Jian, S. Y. B.; Nayak, T.; Majumder, N.; and Poria, S. 2021. Aspect sentiment triplet extraction using reinforcement learning. In *CIKM 2021*, 3603–3607.

Jøsang, A. 2016. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer.

Khattak, A.; Habib, A.; Asghar, M. Z.; Subhan, F.; et al. 2021. Applying deep neural networks for user intention identification. *Soft Computing.*, 25(3): 2191–2220.

Koohikamali, M.; and Sidorova, A. 2017. Information Re-Sharing on Social Network Sites in the Age of Fake News. *Informing Science*, 20.

Purohit, H.; and Pandey, R. 2019. Intent mining for the good, bad, and ugly use of social web: Concepts, methods, and challenges. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, 3–18. Springer.

Shen, Y.-C.; Lee, C. T.; Pan, L.-Y.; and Lee, C.-Y. 2021. Why people spread rumors on social media: Developing and validating a multi-attribute model of online rumor dissemination. *Online Information Review*.

Shu, K.; Mukherjee, S.; Zheng, G.; et al. 2020. Learning with weak supervision for email intent detection. In *ACM SIGIR 2020*, 1051–1060.

Wang, T.; Zhou, J.; Hu, Q. V.; and He, L. 2019. Aspect-level sentiment classification with reinforcement learning. In *2019 IJCNN*, 1–8.

Wang, W. Y. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *ACL 2017 (Volume 2: Short Papers)*, 422–426.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3): 229–256.

Yogatama, D.; Blunsom, P.; Dyer, C.; Grefenstette, E.; and Ling, W. 2017. Learning to compose words into sentences with reinforcement learning. In *ICLR 2017*.

Zhang, T.; Huang, M.; and Zhao, L. 2018. Learning structured representation for text classification via reinforcement learning. In *Thirty-Second AAAI Conference*.