Noname manuscript No.
(will be inserted by the editor)

LENS: Label Sparsity-Tolerant Adversarial Learning on Spatial Deceptive Reviews

Sirish Prabakar · Haiquan Chen (corresponding author) · Zhe Jiang · Carl Yang · Weikuan Yu · Da Yan

Received: date / Accepted: date

Abstract Online businesses and websites have recently become the main target of fake reviews, where fake reviews are intentionally composed to manipulate the business ratings positively or negatively. Most of existing works to detect fake reviews are supervised methods, whose performance highly depends on the amount, quality, and variety of the labeled data, which are often non-trivial to obtain in practice. In this paper, we propose a semi-supervised label sparsity-tolerant framework, LENS, for fake review detection by mining spatial knowledge and learning distributions of embedded topics. LENS builds on two key observations. (1) Spatial knowledge revealed in spatial entities and their co-occurring latent topic distributions may indicate the review authenticity. (2) Distributions of the embedded topics (the contextual distribution) may exhibit important patterns to differentiate between real and fake reviews. Specifically, LENS first extracts embeddings for spatial named entities using a knowledge base trained from Wikipedia webpages. Second, LENS represents each input token as a distribution over the learned latent topics in the embedded topic space. To bypass the differentiation

Sirish Prabakar
California State University, Sacramento, CA, USA
E-mail: sirishprabakar@csus.edu

Haiquan Chen
California State University, Sacramento, CA, USA
E-mail: haiquan.chen@csus.edu

Zhe Jiang
University of Florida, Gainesville, FL, USA
E-mail: zhe.jiang@ufl.edu

Carl Yang
Emory University, Atlanta, GA, USA
E-mail: j.carlyang@emory.edu

Weikuan Yu
Florida State University, Tallahassee, FL, USA
E-mail: yuw@cs.fsu.edu

Da Yan
Indiana University, Bloomington, IN, USA
E-mail: yanda@iu.edu

difficulty, LENS builds on two discriminators in the actor-critic architecture using reinforcement learning. Extensive experiments using the real-world spatial and non-spatial datasets show that LENS consistently outperformed the state-of-the-art semi-supervised fake review detection methods on few labels at all different labeling rates for real and fake reviews, respectively, in a label-starving setting.

Keywords Fake review detection, Label sparsity, Reinforcement learning, Generative adversarial networks

1 Introduction

Recently fake reviews have been widespread on online review websites and have obtained significant research attention. Fake reviews can either aim to promote a business or tarnish the reputation of rivalry businesses by manipulating the overall perception of a service or a product. However, most of existing solutions [8, 28–30, 36] to detecting online fake reviews build on supervised learning methods, which rely on learning the effective lexical and syntactic patterns from a large amount of labeled reviews as ground truth. Therefore, the performance of those supervised learning methods highly depends on the amount, quality, and variety of the labeled data, which are often non-trivial to obtain in practice. It is desirable to have an effective fake review detection approach that is able to combat the label sparsity issue by accurately distinguish between genuine (real) reviews and deceptive (fake) reviews while only requiring a small amount of labeled training data.

Generative Adversarial Network (GAN) based methods [9, 33, 35] have recently been proposed for text generation as semi-supervised learning techniques. A GAN framework consists of two models: a generative model which tries to learn the data distribution and a discriminator that classifies whether a sample comes from the training data or from the generator. Such GAN framework simulates a minimax two-player game. Two GAN-based semi-supervised approaches, FakeGAN [1] and SpamGAN [31], have been proposed for online fake review detection, which have showed promising results when handling limited labeled training data. On the other hand, as a general semi-supervised solution for text classification, GAN-BERT [4] integrated a pre-trained BERT [5] encoder with a generative adversarial network by jointly learning from labelled and unlabeled data to alleviate the label-starving problem. Recently, CEST [32] has been proposed as a new semi-supervised framework for text classification on few labels. CEST employs BERT as the encoder and constructs a contrast-enhanced similarity graph to utilize data efficiently. However, all the aforementioned semi-supervised approaches for fake review detection or text classification suffer from two major limitations: (1) They do not distinguish between spatial reviews and non-spatial reviews, therefore ignoring the spatial knowledge that can be potentially leveraged to enhance the inference results. (2) They generate the synthetic reviews in the latent neural word embedding space at the word (token) level and therefore fail to consider the important distribution patterns among the topics discussed in those reviews.

Our Observations and Contributions. Motivated by the two aforementioned drawbacks of existing approaches for fake review detection, in this paper, we propose a semi-supervised label sparsity-tolerant framework, LENS (Learning Embedded coNtextual DiStribution), for fake review detection by mining external spatial knowledge and learning the distribution patterns of the latent embedded topics. LENS builds on two of our key observations. (1) *Spatial knowledge revealed in spatial entities and their co-occurring latent topic distributions may indicate the review authenticity.* (2) *Distributions of the embedded topics [6] (the contextual distribution) may exhibit important patterns to differentiate between real and fake reviews*. We summarize our contributions as follows:

- We propose LENS, a semi-supervised label sparsity-tolerant framework for fake review detection by mining spatial knowledge from spatial named entities and learning the distribution of the latent embedded topics. Unlike existing works on fake review detection methods with few labels [1,4,23,31,32], LENS builds on topic-space representations by learning from the global semantics (topic distribution) rather than using the local semantics (word embedding).
- To capture the spatial knowledge revealed in the reviews, we extract the spatial named entities and obtain their latent representations by learning from a knowledge base trained with Wikipedia webpages.
- To learn the important patterns exhibited in the latent embedded topics in reviews, we represent each input token (word or spatial named entity) as a distribution over the learned embedded topics, i.e., the contextual distribution.
- To tackle the mode collapse issue [2,12,21], LENS builds on the actor-critic architecture with two discriminators using policy gradient in reinforcement learning. Specifically, one discriminator differentiates between real and fake reviews while the other discriminator differentiates between the fake reviews from the dataset and the fake reviews from the generator.
- Extensive experiments on Yelp-based spatial and non-spatial datasets show that LENS consistently outperformed the state-of-the-art semi-supervised fake review detection methods on few labels at all different labeling rates for real and fake reviews, with up to 27% and 31% improvements on accuracy and F1-score, respectively.

The remaining sections are organized as follows: Section 2 introduces the datasets created in our study. Section 3 elaborate on our key observations while Section 4 discusses LENS architecture. Embedding extraction of spatial named entities using knowledge base is discussed in Section 5. Section 6 discusses embedded topic modeling while Section 7 discusses the dual discriminator architecture. Comparative experiment results and a case study are presented in Section 8. Section 9 reviews the related work while Section 10 concludes the paper.

2 Creation of the Benchmark Datasets

YelpZip. Although opinion spamming has been widespread, there are not many commercial websites that filter fake reviews. Yelp[1] implements review filtering on a commercial scale. The filtered reviews on Yelp for each business can be accessed through a link at the bottom of the Yelp page of that business. While the Yelp filtering mechanism is not perfect, it has been proven in the literature to produce accurate results [26,27]. Therefore, in our study, we used the YelpZip [26,27] dataset as ground truth, which treated the recommended and filtered reviews on

[1] https://www.yelp.com

Table 1: Basic statistics of the YelpZip, Yelp-Spatial, and Yelp-Non-Spatial datasets.

Dataset	# of total reviews	# of real reviews (%)	# of fake reviews (%)
YelpZip	608,598	528,142 (87%)	80,456 (13%)
Yelp-Spatial	223,954	197,946 (88%)	26,008 (12%)
Yelp-Non-Spatial	367,303	314,449 (86%)	52,854 (14%)

Table 2: Four particular types of spatial named entities extracted to construct our Yelp-Spatial dataset.

Type	Description
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states, etc.
LOC	mountain ranges, bodies of water, etc.

Table 3: Samples of unedited real and fake reviews in YelpZip (the extracted spatial named entities are in bold).

Unedited Yelp Reviews	
Real: Spatial:	i to **italy** travel a lot and am a confirmed pizza-snob, but after hearing about this place had high expectations. every bit as good as the pizza i have enjoyed in **napoli**. it was well-worth the short detour off of i-7, even worth a long detour!
Real: Non-Spatial:	absolutely adorable. unique sandwiches. great atmosphere. wonderful staff. you can't go wrong with cheeky sandwiches. keep your eyes peeled when looking for it though, anywhere in town.
Fake: Spatial:	i love this bar. the food, the staff, the fish, it's all good. i like that the music is always cool, but different depending on which bartender is working . i also really like the french fries the backyard is one of the most pleasant in **brooklyn**.
Fake: Non-spatial:	we've now made this our ”sunday night spot” and can't picture a better place to eat, share a bottle of wine, and relax. the service is exceptional, and we never feel rushed or pushed to leave. it's a very laid back place with an awesome atmosphere.

Yelp as real and fake, respectively. Concretely, YelpZip was crawled from the Yelp web pages by looping over zipcode numbers incrementally. The process started with a zipcode in NY state, collected all the reviews for restaurants in that zipcode, and increased the zipcode number incrementally. All the zipcodes were organized by geography in order to retrieve unbiased reviews for restaurants in a continuous region on the U.S. map. The summary statistics of the YelpZip dataset are given in Table 1. We observe that many of the reviews in YelpZip contain location-related information. Therefore, we used Named Entity Recognition (NER)[2] to identify and extract the named entities related to locations, organization, and places. We removed all the trivial reviews with no more than 5 words and divided the remaining reviews into two subsets: Yelp-Spatial and Yelp-Non-Spatial.

[2] https://spacy.io/

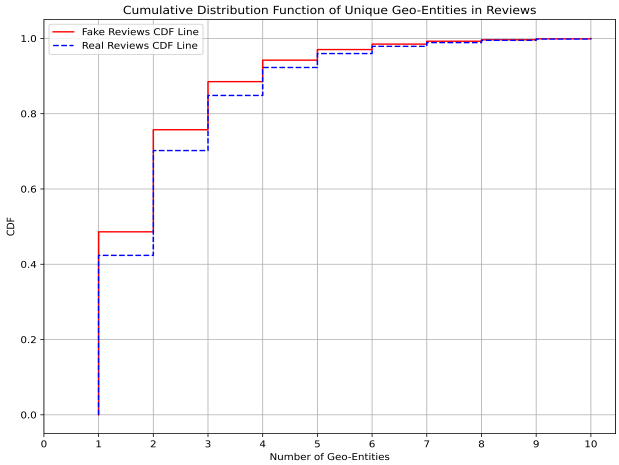


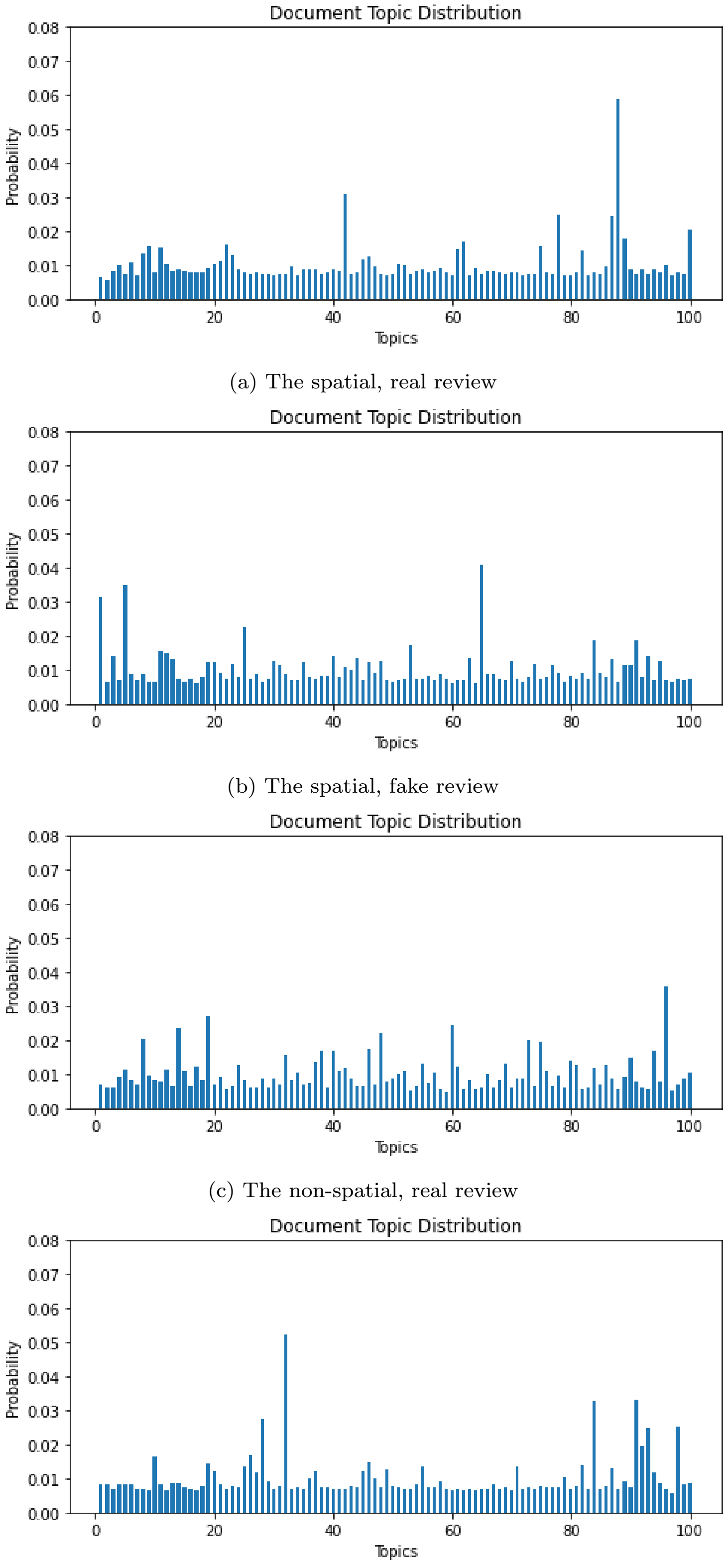
Fig. 1: Cumulative distribution functions of the number of unique spatial named entities (removing duplicates) mentioned in each real review and each fake review, respectively, in Yelp-Spatial.

Yelp-Spatial. This version of YelpZip contains all the reviews with one or more than one spatial named entity in one of the four following categories: LOC, GPE, FAC, and ORG. The description of each exacted category is as shown in Table 2. The basic statistics of Yelp-Spatial are in Table 1.

Yelp-Non-Spatial. This version of YelpZip contains all the reviews without any spatial named entity in any of the aforementioned four categories. The basic statistics of Yelp-Non-Spatial are in Table 1.

3 Key Observations

Observation 1: Spatial knowledge revealed in spatial entities and their co-occurring latent topic distributions may indicate the review authenticity. Figure 1 depicts the cumulative distribution functions of the number of unique spatial entities (removing duplicates) mentioned in each real review and each fake review in Yelp-Spatial. Specifically, as shown in Figure 1, 70% of real reviews mention one or two unique spatial entities while 77% of fake reviews mention one or two unique spatial entities. Table 3 presents four random unedited reviews in our created datasets, based on whether the review has spatial information and its authenticity. For example, the first review is a real, spatial review mentioning "*italy*" and "*napoli*" while the third review is a fake, spatial review mentioning "*brooklyn*". We observe that by only examining the review mentioning "*italy*", "*napoli*", or "*brooklyn*" itself, it would be difficult to infer their authenticity. However, if the latent topic distribution exhibited in a review mentioning some spatial entities was significantly different from those exhibited in other reviews mention-



(a) The spatial, real review

(b) The spatial, fake review

(c) The non-spatial, real review

(d) The non-spatial, fake review

Fig. 2: Distribution over the learned latent embedded topics (100 topics learned on the YelpZip dataset) for each review as shown in Table 3

ing the same spatial entities, such topic-level discrepancy could be leveraged to infer authenticity or detect fakeness. This motivates us to detect fake reviews by mining the spatial knowledge revealed in the spatial entities and exploiting the latent topic distributions co-occurring with those spatial entities.

Observation 2: Distributions of the embedded topics [6] (the contextual distribution) may exhibit important patterns to differentiate between real and fake reviews. The word embedding techniques commonly used in the literatures (e.g., word2vec, GloVe, and BERT) build on local co-occurrences and local semantics, which usually lead to very high-dimensional and sparse representations. We observe that the topic-space representations (topic distributions) based on the topics (global semantics) discovered from entire collections of documents can leveraged to detect fakeness. We extracted the embedded topics [6] by combining the Latent Dirichlet Allocation (LDA) with the neural word embeddings [3]. We then visualized the learned embedded topic distributions of four random selected reviews in Table 3 by learning 100 embedded topics on the YelpZip dataset. Figure 2(a) and Figure 2(b) show the topic distributions of the 100 discovered topics for the spatial, real review and the spatial, fake review, respectively, while Figure 2(c) and Figure 2(d) show the topic distributions of the 100 discovered topics for the non-spatial, real review and the non-spatial, fake review, respectively. As shown in Figure 2, the reviews in all four different categories show unique patterns in their respective topic distributions. Specifically, Our observations are twofold. First, as depicted in Figure 2(a) and Figure 2(b), the two spatial reviews (reviews mentioning at least one spatial named entity) exhibit different embedded topic distributions. Second, regardless of whether a review mentions any spatial entities or not, the real and fake reviews always exhibit different embedded topic distributions. Note that our goal is not to directly quantify the difference between the embedded topic distributions in real and fake reviews, but to train a generative model to learn and generate such discrepancy in the embedded topic space, which can be effectively leveraged to enhance the detection of review fakeness in a label-starving setting. Specifically, training a generative model in the embedded topic space yields two major benefits. First, using topic-space embeddings encodes the reviews with global semantics (compared to word embeddings which builds on local co-occurrences and local semantics), which improves the training accuracy. Second, using topic-space embeddings reduces the dimensionality of the training data (i.e., we used 100 topics by default, compared to 300 used by word2vec and GloVe, and 768 used by BERT), which improves the training efficiency.

4 LENS Overview.

LENS Architecture. Figure 3 shows the architecture of LENS. Note that unlike existing works on semi-supervised text classification or fake review detection [1,4,23,31,32], LENS takes advantage of topic-level representations by learning from the global semantics (topic distribution) rather than using the local semantics (word embedding). LENS consists of three modules: (1) extracting embedding of spatial entities using knowledge base, (2) representing contextual distribution using embedded topics, (3) generating and classifying fake reviews using GAN in the actor-critic architecture. First, LENS extracts spatial named entities using knowledge base to obtain their embeddings trained from Wikipedia webpages. Second,

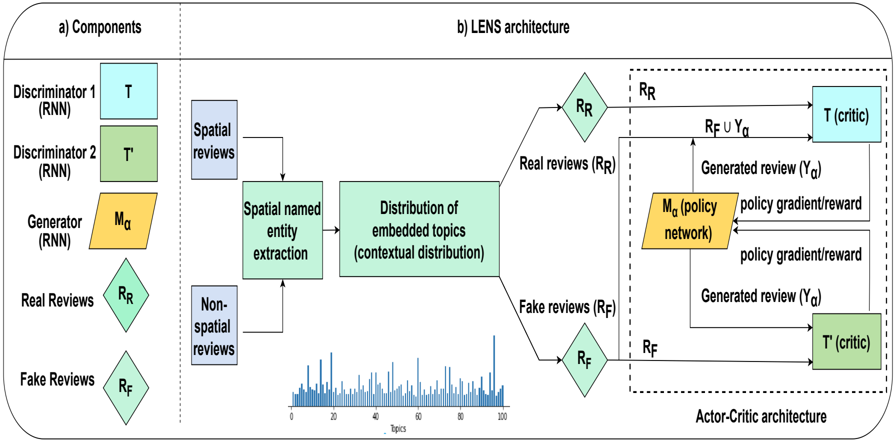


Fig. 3: (a) LENS Components. (b) LENS architecture.

LENS represents each input token (word or named entity) as a distribution over the learned latent topics in the embedded topic space. To tackle the mode collapse issue [2, 12, 21], LENS builds on the actor-critic architecture with two discriminators using policy gradient in reinforcement learning. Specifically, one discriminator differentiates between real and fake reviews while the other discriminator differentiates between the fake reviews from the dataset and the fake reviews from the generator. As illustrated in Figure 3, the generator M_α learns the data distribution from both the real and fake reviews. The discriminator T tries to distinguish real reviews from fake reviews or the real-looking reviews from the generator while the discriminator T' tries to distinguish fake reviews and the fake-looking reviews from the generator. Both discriminators are trained together during the training stage with the generator trying to fool the two discriminators. On the other hand, the generator in LENS is trained as policy (actor) via policy gradient [35]. The reward signal for each complete sequence (review) is provided by dual discriminators and Monte Carlo tree search is used to approximate the state-action value for gradient update in policy. By receiving the feedback from both discriminators, the generator updates its parameters to produce more realistic fake reviews, which in turn help training stronger discriminators.

5 Embedding Extraction for Spatial Named Entities using Knowledge Base

In this section, we discuss on embedding extraction for spatial named entities using a knowledge base. We first extracted all spatial named entities using wiki2vec [34], which was pre-trained using Wikipedia pages to represent each named entity as its embedding. Table 4 shows some examples of spatial named entities extracted

Table 4: Sample spatial named entities in Yelp-Spatial included in wiki2vec (in the original wiki2vec entity format)

Spatial Named Entities
ENTITY/The_Ny_Botanical_Garden
ENTITY/The_Brooklyn_Bridge
ENTITY/Iron_Chef_House
ENTITY/East_Passyunk_Ave
ENTITY/The_Woodbridge_Center_Mall
ENTITY/Lutheran_Medical_Center
ENTITY/The_Reading_Terminal_Market
ENTITY/The_U_of_Penn_Hospital

Table 5: Examples of the fuzzy-matching based closest matches for the words/spatial named entities that are not included in wiki2vec. A higher score represents a higher similarity.

Unrecognized word/entity	Closest matching words/entities
”chowed” **(word)**	(’chewed’, 91), (’showed’, 83), (’choked’, 83), (’chow-mein’, 83), (’echoed’, 83)
”underseasoned” **(word)**	(’unseasoned’, 87), (’unreasoned’, 87), (’underseas’, 82), (’underspanned’, 80), (’understand’, 78)
”bloomingdale” **(spatial entity)**	(”Bloomingdale’s”, 100), (’Bloomingdale,_Illinois’, 100), (’Bloomingdale,_New_Jersey’, 100)
”SriPraPhai” **(spatial entity)**	(’praphai’, 82) , (’sasiprapha’, 80) , (’sérigraphie’, 80) , (’sriburapha’, 80) , (’sriracha’, 78)

using wiki2vec. Since online reviews abound with abbreviations, slang words, and typos, for each word and spatial named entity not found in the knowledge base, we calculated its edit distance (Levenshtein distance) using fuzzy string matching to identify the most matching word or spatial named entity. Specifically, we assign each unrecognized word or spatial entity in the reviews to the closest matching word or spatial entity in our wiki2vec-based knowledge base. Table 5 shows four examples of the words and extracted spatial named entities that are not found in the wiki2vec corpus and their most similar counterparts in the wiki2vec corpus. A higher score represents a higher similarity.

6 Topic distribution representation using embedded topics

Embedded topic extraction. LENS extracts embedded topics [6] by combining LDA with the neural word embeddings [3]. Embedded topics have been proven to be able to provide more meaningful and interpretable topics and yield better performance in terms of predictive accuracy than LDA. Specifically, by learning embedded topics, LENS treats each word as a categorical distribution whose natu-

Algorithm 1 Learning embedded topic embeddings in LENS

```
Initialize model and variational parameters.
foreach $i$ do
    Compute $\beta_k = \mathrm{softmax}\left(\rho^\top \alpha_k\right)$ for each topic $k$
    Choose a minibatch $\mathbb{B}$ of documents
    foreach document $d \in \mathbb{B}$ do
        Get word embedding representation $\mathbf{x}_d$
        Compute $\mu_d = \mathrm{NN}\left(\mathbf{x}_d; \nu_\mu\right)$
        Compute $\Sigma_d = \mathrm{NN}\left(\mathbf{x}_d; \nu_\Sigma\right)$
        Sample $\theta_d \sim \mathcal{LN}\left(\mu_d, \Sigma_d\right)$
        foreach word $w_{dn} \in d$ do
            Compute $p\left(w_{dn} \mid \theta_d\right) = \theta_d^\top \beta_{\cdot, w_{dn}}$
        end
    end
    Estimate the evidence lower bound (ELBO) and its gradient using backpropagation
    Update model parameters $\alpha_{1:K}$
    Update variational parameters $(\nu_\mu, \nu_\Sigma)$
end
```

ral parameter is the inner product between its word embedding and the embedding of its assigned topic. The higher the inner product is, the more likely a word belongs to a particular topic. Let ρ be an $L \times V$ matrix containing L-dimensional embeddings of the words in the vocabulary and the extracted spatial named entities, i.e., each column $\rho_k \in \mathbb{R}^L$ corresponds to the word/wiki2vec embedding representation of the k^{th} word/spatial named entity. We can define each topic β_k (the word distribution of each discovered topic) as follows

$$\beta_k = \mathrm{softmax}\left(\rho^\top \alpha_k\right), \tag{1}$$

where $\alpha_k \in \mathbb{R}^L$ is an embedding representation of the k^{th} topic, i.e, the topic embedding. The algorithmic description of extracting topic embeddings is shown in Algorithm 1, where $\mathrm{NN}\left(\mathbf{x}; \nu\right)$ denotes a neural network with input $\mathbf{x}$ and parameters ν, and $\mathcal{LN}\left(\cdot\right)$ denotes the logistic-normal distribution.

Representing each token as its contextual (topical) distribution. Based on the extracted topics, LENS represents the k^{th} word/spatial named entity in the corpus as its corresponding distribution $\mathcal{D}_k$ over all the discovered embedded topics, which can be calculated as

$$\mathcal{D}_k = \mathrm{softmax}\left(\alpha^\top \rho_k\right), \tag{2}$$

where α is a matrix that contains the topic embedding of each topic and ρ_k is the word/wiki2vec embedding of the k^{th} word/spatial named entity. Instead of directly using the word embeddings ρ_k, LENS converts each review $\mathbf{X}_{1:N}$ with N words/named entities to a sequence of word-level topic distributions $\{\mathcal{D}_1, \mathcal{D}_2, ...\mathcal{D}_N\}$, which is then fed as input to the subsequent review generation and classification module.

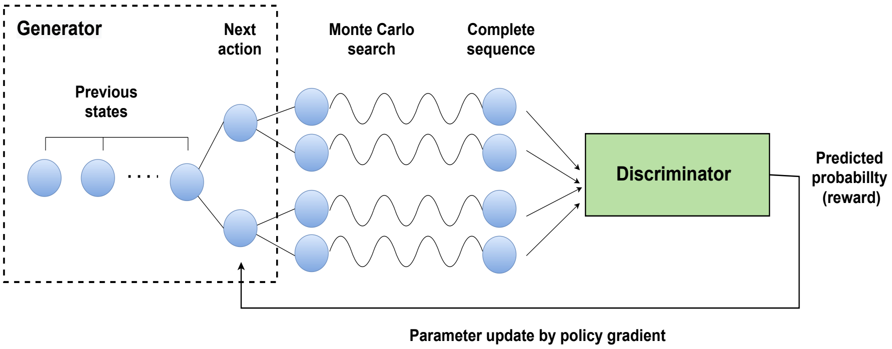


Fig. 4: LENS generator is trained via policy gradient by approximating an intermediate state-action value using Monte Carlo tree search, where the probability (reward) for a final outcome (a complete review) to be real/fake is predicted by the two discriminators .

7 Review generation and classification using the actor-critic architecture

Dual discriminators. To bypass the differentiation difficulty due to the discrete output from the generator [35], LENS builds on the actor-critic architecture in reinforcement learning. To tackle the mode collapse issue [2, 12, 21], LENS employs two discriminators (critics) to learn from both the real and fake review distributions. Suppose the two discriminator models are denoted as T and T'. We represent each review as $\mathrm{P}_{1:L}$, which has L tokens. All the reviews created by the generator M_α is represented as Y_G. The discriminator T differentiates between real and fake reviews by predicting $\mathrm{T}(\mathrm{P}_{1:L})$, which is the probability indicating whether the review $\mathrm{P}_{1:L}$ belongs to R_R (real reviews) or R_F (fake review) $\cup$ Y_G (synthetic reviews). On the other hand, the discriminator T' distinguishes between fake reviews R_F and the reviews generated by M_α by predicting $\mathrm{T}'(\mathrm{P}_{1:L})$, which is the probability indicating whether the review belongs to R_F or Y_G. Both discriminators were trained as a binary classifier with the sigmoid cross-entropy as the loss function. The two discriminators adopt the same LSTM [13] auto-regressive sequence-to-sequence architecture, followed by a fully connected layer with the sigmoid activation function.

Generator. Trained via policy gradient [35], the generator M_α functions as a policy network (actor) in the actor-critic architecture by generating a review token by token. The reward signal for each complete sequence (review) is provided by dual discriminators and Monte Carlo tree search is used to approximate the state-action value for gradient update in policy as shown in Figure 4. At step t, the state s is the sequence of all the tokens generated so far and the action a is the next word. Both discriminators provide the reward for a complete sequence (a complete review), which is the predicted probability for the review to be real. Note that the generator M_α receives rewards from both the discriminators. Each type of the feedback help improving the quality of the reviews generated by the generator in its own way. Concretely, the discriminator T' provides the feedback to help the generator produce the reviews as close to the reviews in R_F as possible. Similarly, the discriminator T provides the feedback to help the generator produce the reviews as close to the reviews in R_R as possible. As the generator obtains feedback from both T and T', the generator M_α is trained to fool the discriminator T' by generating the reviews that seems fake and fool the discriminator T by generating the reviews that seems real. The generator in LENS adopts the LSTM [13] auto-regressive sequence-to-sequence architecture, followed by the softmax activation function.

8 EXPERIMENTAL VALIDATION.

In this section, we empirically show the superiority of LENS over the state of the art. We begin with the experimental setting and the methods to compare with, followed by their performance comparison. The parameter sensitivity analysis and a case study were also provided.

8.1 Experimental Setting

In order to evaluate the model performance in the label-starving setting, we randomly selected 400 reviews (200 true reviews and 200 fake reviews) from Yelp-Spatial and Yelp-Non-Spatial datasets, respectively. We divided the reviews into training and test sets with a split of 80% and 20%. Note that we follow the same experimental setting as used in the state-of-the-art semi-supervised text classification approaches on few labels [1, 4, 31, 32] by varying the percentage of the labeled records in the training set (labeling rate) from 10%, 30%, 50%, 70%, to 90% (which correspond to 16, 48, 80, 112, and 144 labels per class in our training set) to investigate the impact of label sparsity on the model performance. All the experiments were conducted on a machine of Tesla P100 GPU and 25 GB RAM with Python v3.7 and TensorFlow v1.15.0 installed. Each result reported in this section was averaged over three runs.

8.2 Methods of Comparison

We obtained the official code of all the baseline methods from their GitHub pages and implemented the following methods for performance comparison.

- **SpamGAN [31]:** This semi-supervised method builds on generative adversarial network with an auxiliary classifier (ACGAN) [23] to detect online fake reviews with limited labeled data. It models the sequence generation as an reinforcement learning problem to handle longer sentences.
- **FakeGAN [1]:** This semi-supervised method uses two discriminators to learn from the real and fake reviews and one generator as policy network to receive rewards from both discriminators for review generation.
- **GAN-BERT [4]:** This semi-supervised method fine-tunes a BERT [5]-based pre-trained encoder under the framework of a generative adversarial network to jointly learn from labeled and unlabeled data.

- **CEST [32]**: This semi-supervised method employs BERT [5] as the encoder and constructs a contrast-enhanced similarity graph with a Bayesian neural network to provide better certainty estimates for unlabeled data to improve the accuracy of pseudo labels during self training.
- **LENS-E (Entity extraction)):** This is the version of LENS with only the extraction of spatial named entities, where the embedding of each word and the embedding of each extracted spatial named entity (trained via the respective Wikipedia pages) are sent to the review generation and classification module. Neither fuzzy matching nor embedded topic learning is used.
- **LENS-F (Fuzzy matching):** This is the version of LENS with the extraction of spatial named entities and fuzzy matching, where each unrecognized word and spatial named entity is mapped to their closest counterparts using fuzzy matching. No embedded topic learning is used.
- **LENS-ETM (Embedded Topic Modeling):** This is the version of LENS with the extraction of spatial named entities and the embedded topics, where the contextual (topical) distribution of each word and each spatial named entity is fed to the review generation and classification module. No fuzzy matching is used.
- **LENS:** This is the complete version of LENS. The two discriminators and the generator were implemented using the same LSTM [13] auto-regressive sequence-to-sequence architecture (with a latent embedding size of 64 and a maximum sequence length of 200). We set the generator to always generate the reviews with a fixed length of 200 words and applied zero padding to the reviews with less than 200 words during training. The generator and the two discriminators were trained using an Adam optimizer with a learning rate of 0.01 and 0.0001, respectively, and a batch size of 64. To balance the generator and the discriminators, we updated the two discriminators 5 times per generator update. We used the pretrained GloVe model for word embeddings. The default word embedding size and the default number of embedded topics to learn in LENS were 100.

8.3 Performance Comparison and Ablation Study

Performance on spatial data. Here we compare the performance of all the methods in terms of accuracy and F1-score on spatial data by varying the percentage of the labeled data in the training set. As shown in Table 6 and Table 7, LENS outperformed the state-of-the-art methods and all its variants for both the real and fake reviews, respectively. When we increased the percentage of the labeled data in the training set, both the accuracy and F1-score improved accordingly. Specifically, when we used 50% of the labeled training data, SpamGAN, FakeGAN, GAN-BERT, and LENS returned a F1-score of 64.56%, 70.06%, 68.32%, and 71.84%, respectively, for real reviews while they yielded a F1-score of 52.46%, 33.82%, 34.80%, and 62.22%, respectively, for fake reviews.

Performance on non-spatial data. Here we compare the performance of all the methods in terms of accuracy and F1-score on non-spatial data by varying the percentage of the labeled data in the training set. As shown in Table 8 and Table 9, LENS outperformed the state-of-the-art methods and all its variants for both the real and fake reviews, respectively. When we raised the percentage of the labeled data in the training set, both the accuracy and F1-score increased accordingly. Specifically, when we used 50% of the labeled training data, Spam-GAN, FakeGAN, GAN-BERT, and LENS returned a F1-score of 71.33%, 62.92%, 67.51%, and 72.46%, respectively, for real reviews while they yielded a F1-score of 62.22%, 38.88%, 60.30%, 62.87%, respectively, for fake reviews.

Performance analysis. The performance gain of LENS over all the baseline methods in a label-starving setting can be attributed to three major features in LENS. First, LENS extracts and leverages external spatial knowledge revealed in spatial named entities and their co-occurring latent topic distributions to tackle label sparsity. Second, LENS generates synthetic reviews in the pre-trained embedded topic space, instead of in the pre-trained word embedding space. Compared to the word embeddings, which build on local co-occurrences and local semantics, topic-space embeddings encode the reviews with rich, global semantics, which improves training accuracy. Also, using topic-space embeddings reduces the dimensionality of the training data (i.e., we used 100 topics by default, compared to 300 used by word2vec and GloVe, and 768 used by BERT), which improves the training efficiency. Third, the generative framework with dual discriminators in LENS forces the model to generate both high-quality real-looking and fake-looking synthetic reviews simultaneously, which augments the supervisory signals/labels that can be used for classification model training in a label-starving setting.

8.4 Parameter Sensitivity Analysis

In this section, we varied the number of embedded topics to investigate its impact on the performance of LENS-ETM and LENS. Given the number of topics, we gradually increased the percentage of the labeled data in the training set from 10% to 100% and reported the average results for that particular number of topics.

Impact of the number of topics on spatial data. Tables 10 and Table 11 show impact of the number of topics on the performance of LENS-ETM and LENS on spatial data for real and fake reviews, respectively, when we varied the number of topics from 10, 50, 100, 150 to 200. Under all the settings, LENS successfully yielded an accuracy and F1-score of more than 61%, which verifies the robustness of LENS. It can also be observed that the optimal number of topics for our Yelp-Spatial data is 100 for both real and fake reviews, with F1-score of 80.59% and 62.79%, respectively.

Impact of the number of topics on non-spatial data. Table 12 and Table 13 show the impact of the number of topics on the performance of LENS-ETM and LENS on non-spatial data for real and fake reviews, respectively, when we varied the number of topics from 10, 50, 100, 150 to 200. Under all the settings, LENS successfully yielded an accuracy and F1-score of higher than 56%, which verifies the robustness of LENS. It can also be observed that the optimal number of topics for our Yelp-Spatial data is 100 for both real and fake reviews, with a F1-score of 66.44% and 58.95%, respectively.

Table 6: Performance comparison on spatial data (Yelp-Spatial) for real reviews.

Labeling rate (# of training labels per class)	10% (16)		30% (32)		50% (180)		70% (112)		90% (144)	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SpamGAN	66.66	67.98	71.60	66.81	69.45	64.56	70.36	65.29	74.06	65.43
FakeGAN	60.46	69.33	59.13	71.89	59.51	70.06	59.57	71.33	58.76	70.37
GAN-BERT	57.39	62.31	66.72	69.28	65.56	68.32	71.28	70.81	71.88	72.85
CEST	61.41	66.05	68.84	70.74	70.02	70.97	72.41	72.23	72.36	73.34
LENS-E	61.62	70.19	59.77	69.33	62.79	71.48	58.71	70.32	63.91	73.82
LENS-F	65.88	69.17	69.31	68.63	70.05	70.73	72.21	72.44	72.92	73.85
LENS-ETM	66.66	68.14	69.11	70.14	69.69	69.31	71.06	66.09	69.01	71.53
LENS	**71.50**	**71.87**	**73.75**	**73.67**	**73.71**	**73.84**	**73.75**	**73.56**	**74.35**	**74.76**

Table 7: Performance comparison on spatial data (Yelp-Spatial) for fake reviews.

Labeling rate (# of training labels per class)	10% (16)		30% (48)		50% (80)		70% (112)		90% (144)	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SpamGAN	53.70	50.04	48.76	51.27	46.91	52.46	54.31	54.26	47.52	45.34
FakeGAN	50.24	34.28	57.14	35.82	60.20	33.82	65.30	36.61	65.21	38.46
GAN-BERT	41.15	32.31	51.41	33.57	59.65	34.80	67.80	39.65	68.65	59.89
CEST	54.85	55.22	65.52	55.92	69.65	57.58	72.55	63.22	74.14	66.63
LENS-E	58.62	43.03	59.25	41.02	67.85	48.10	75.67	50.90	71.42	50.63
LENS-F	67.74	51.85	69.58	53.65	71.65	57.83	75.43	64.41	76.33	60.24
LENS-ETM	59.18	57.42	62.14	60.78	68.78	60.28	59.87	64.13	63.82	60.60
LENS	**68.42**	**57.77**	**71.05**	**61.67**	**73.68**	**62.22**	**77.50**	**67.39**	**78.22**	**68.08**

Table 8: Performance comparison on non-spatial data (Yelp-Non-Spatial) for real reviews.

Labeling rate (# of training labels per class)	10% (16)		30% (48)		50% (80)		70% (112)		90% (144)	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SpamGAN	67.50	69.97	66.75	70.34	67.71	71.33	68.75	73.72	70.35	74.15
FakeGAN	51.54	59.52	53.39	61.97	54.36	62.92	54.00	62.42	54.73	62.65
GAN-BERT	52.45	60.75	55.72	65.78	63.90	67.51	63.82	66.54	71.40	74.37
CEST	68.17	70.61	68.10	71.70	68.37	71.56	68.61	72.50	72.97	74.06
LENS-E	58.82	67.56	60.00	69.38	59.52	68.02	61.17	70.26	62.35	71.62
LENS-F	59.55	69.28	60.19	70.19	61.20	70.27	63.80	62.91	62.79	72.00
LENS-ETM	60.27	63.3	61.42	63.23	62.16	65.71	62.95	55.93	65.75	69.06
LENS	**68.91**	**72.85**	**68.49**	**71.94**	**69.44**	**72.46**	**69.36**	**74.85**	**73.41**	**75.46**

Table 9: Performance comparison on non-spatial data (Yelp-Non-Spatial) for fake reviews.

Labeling rate (# of training labels per class)	10% (16)		30% (48)		50% (80)		70% (112)		90% (144)	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SpamGAN	62.42	57.77	63.05	59.19	64.68	62.22	67.50	61.39	71.19	63.08
FakeGAN	52.27	40.35	50.00	37.03	52.50	38.88	55.81	42.47	57.77	45.61
GAN-BERT	49.08	38.53	51.60	58.52	57.81	60.30	57.61	60.65	71.72	63.11
CEST	62.50	58.61	63.21	58.50	63.59	62.73	67.61	62.90	71.92	63.05
LENS-E	55.17	40.00	59.25	41.55	56.66	41.97	62.06	44.99	65.51	47.50
LENS-F	62.06	43.37	63.47	40.44	62.50	47.61	63.06	53.92	68.75	56.16
LENS-ETM	51.11	47.42	52.08	50.01	54.54	50.00	60.49	48.22	60.00	55.67
LENS	**65.05**	**60.41**	**64.44**	**59.79**	**65.21**	**62.87**	**72.28**	**63.42**	**73.18**	**63.27**

Table 10: Impact of the number of topics on spatial data for real reviews

Number of topics	10		50		100		150		200	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LENS-ETM	67.85	67.46	64.06	60.57	72.52	70.01	70.67	71.8	69.21	69.82
LENS	**81.86**	**80.65**	**83.75**	**79.18**	**83.67**	**80.59**	**77.20**	**77.81**	**76.51**	**76.38**

Table 11: Impact of the number of topics on spatial data for fake reviews

Number of topics	10		50		100		150		200	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LENS-ETM	57.91	57.7	56.26	58.65	61.42	60.21	58.67	57.22	58.31	56.88
LENS	**63.46**	**61.41**	**65.27**	**62.38**	**66.13**	**62.79**	**65.66**	**61.05**	**63.77**	**61.46**

Table 12: Impact of the number of topics on non-spatial data for real reviews

Number of topics	10		50		100		150		200	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LENS-ETM	60.38	60.32	61.44	61.12	63.61	60.80	63.08	57.76	64.24	56.02
LENS	**75.3**	**66.83**	**71.56**	**65.42**	**75.28**	**66.44**	**65.78**	**63.7**	**74.2**	**59.73**

Table 13: Impact of the number of topics on non-spatial data for fake reviews

Number of topics	10		50		100		150		200	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LENS-ETM	56.32	55.38	57.04	56.06	59.83	57.21	57.24	55.84	58.68	54.82
LENS	**58.64**	**57.39**	**58.97**	**58.28**	**60.16**	**58.94**	**59.24**	**56.72**	**58.72**	**57.11**

8.5 Case Study on Learned Embedded Topics

Table 14 presents the eight randomly selected topics out of the 100 topics dicsovered using LENS based on our created Yelp-Spatial dataset, along with their top words (words with the closest distances to each topic embedding). Here we labeled each topic with a probable theme it describes. For example, Topic 1 describes the topic "*taste*" while Topic 2 describes the topic "*smell*". As Table 14 shows, the top words in each topic consistently show both the desirable topic coherence and diversity, which verifies the effectiveness of extracting the embedded topics in LENS.

9 Related Work

In this section, we summarize the related work on the unsupervised topic modeling techniques and semi-supervised fake review detection methods. We reveal their major limitations and explain the advantages of LENS over existing methods when tackling the label sparsity issue for fake review detection.

Unsupervised Topic modeling As an unsupervised learning technique to identify latent topics in text documents, topic models are widely used in feature selec-

Table 14: Sample discovered topics using LENS on our created Yelp-Spatial dataset and their corresponding top words. Note that here we labeled each topic with a probable theme it describes. For example, Topic 1 describes ”taste” while Topic 2 describes ”smell”.

Topic 1 [taste]	buds	tasted	tastes	flavor	fishy	texture	salty	aftertaste
Topic 2 [smell]	smells	sewage	smelled	odor	chemical	smelling	putrid	rancid
Topic 3 [price]	prices	reasonable	quality	pricing	quantity	bucks	value	expensive
Topic 4 [location]	building	rittenhouse	parking	spot	blvd	waterfront	west	central
Topic 5 [distance]	blocks	block	within	walking	miles	mile	radius	nearest
Topic 6 [service]	serive	staff	services	ambiance	sevice	atmosphere	courteous	unattentive
Topic 7 [speed]	dial	improving	fast	rapid	slow	piling	disapointing	disorganized
Topic 8 [conjunctions]	about	when	like	there	me	out	their	no

tion and clustering documents to obtain insights from unstructured data. Latent Dirichlet Allocation (LDA) is one of the most commonly used approaches in the literature for topic modeling, which uses a generative process with Dirichlet distributions to generate new documents and assign topics. As an extension of LDA on the neural word embedding [3] space, embedded topics modeling (ETM) [6] was recently proposed to learn topic embeddings in the same embedding space as the word embeddings, which aims to enable the arithmetic operations (e.g., a metric function based on cosine distance) directly over each latent topic and individual word. Such learned embedded topics over documents have been proven to be able to provide more meaningful and interpretable topics [6], yielding the state-of the-art performance in various predictive topic modeling tasks.

Semi-supervised fake review detection with feature handcrafting. Traditionally fake review detection techniques leverage supervised learning techniques. Wu et al. [24] used n-gram features to train a classifier based on SVM and Naive Bayes. Ott et al. [15] used logistic regression with reviewer-centric features. The parse trees based on context free grammar were used in [10, 19, 22]. However, supervised learning approaches usually require substantial labeled training data to obtain a desirable performance in practice. Even though there exist a few repositories for online reviews from multiple resources, most of them are unlabeled due to the difficulty to automate the labeling or the expensive time cost for manual labeling. Therefore, semi-supervised fake review detection approaches [17, 18, 20] have been proposed to tackle this challenge with limited labelled training data. However, the performance of their approaches heavily rely on feature handcrafting in order to create a pre-defined set of features to training their classifiers.

Semi-supervised fake review detection without feature handcrafting. As a semi-supervised model, Generative Adversarial Networks (GANs) were traditionally used for generating continuous data (such as images) [7, 11, 25] rather than discrete data like textual documents. Recently, GANs have shown promising results in text classification tasks due to their capability to generate synthetic text [11]. One challenge of training GANs to generate discrete data (e.g., text) is the differentiation difficulty to pass the gradients to update the generator due to the discrete output from the generator [14]). By casting text generation as a reinforcement learning problem, SeqGAN [35] was able to tackle the differentiation difficulty for gradient update. SeqGAN employed the policy gradient to train the generator and the gradient descent to train the discriminator. Inspired by SeqGAN [35], StepGAN [33] and MaskGAN [9] adopted the actor-critic architecture [16] in reinforcement learning to learn the feedback/rewards. All these GAN-based research works focused on generic sentence generation. To specifically tackle the online opinion spams, SpamGAN [31] extended StepGAN [33] by employing an auxiliary classifier. FakeGAN [1] built on SeqGAN [35] and utilized two discriminators to learn from the real and fake reviews, respectively. The general idea behind SpamGAN [31] and FakeGAN [1] is that the discriminator tries to differentiate fake reviews from real ones while the generator tries to generate synthetic reviews to fool the discriminator. As a general semi-supervised solution for text classification, GAN-BERT [4] integrated a pre-trained BERT [5] encoder with a generative adversarial network by jointly learning from labelled and unlabeled data to alleviate the label-starving problem. Recently, CEST [32] has been proposed as a new semi-supervised framework for text classification on few labels. As a certainty-driven sample selection method, CEST employs BERT as the encoder and constructs a contrast-enhanced similarity graph to utilize data efficiently during self training. In self training, pseudo-labels are generated for unlabeled data, which are then used as new labeled data for training. The basic idea of CEST is to use a Bayesian neural network to provide better certainty estimates for unlabeled data and a contrast-enhanced similarity graph to consider smoothness to improve the accuracy of pseudo-labels.

However, all the aforementioned semi-supervised approaches for fake review detection or text classification suffer from two major limitations. (1) They do not distinguish between spatial reviews and non-spatial reviews, therefore ignoring the spatial knowledge that can be potentially leveraged to detect fake reviews. (2) They generate the synthetic reviews in the latent neural word embedding space at the word (token) level and therefore fail to consider the important distribution patterns in the embedded topics exhibited in the reviews. Motivated by these two major limitations, as a semi-supervised label sparsity-tolerant solution, LENS detects fake reviews by mining spatial knowledge and learning the distributions of embedded topics based on reinforcement learning. LENS builds on two of our key observations. (1) *Spatial knowledge revealed in spatial entities and their co-occurring latent topic distributions may indicate the review authenticity.* (2) *Distributions of the embedded topics [6] (the contextual distribution) may exhibit important patterns to differentiate between real and fake reviews*

10 Conclusion

In this paper, we propose a semi-supervised label sparsity-tolerant framework, LENS, for fake review detection by mining spatial knowledge and learning effective distributions of latent topic embedding. Specifically, LENS first extracts spatial named entities using fuzzy matching to obtain their embeddings trained from their respective Wikipedia pages. Second, LENS represents each input token as a contextual distribution over the learned latent topics in the embedded topic space. To bypass the differentiation difficulty due to the discrete output from the generator, LENS builds on the actor-critic architecture in reinforcement learning with two discriminators. Extensive experiments using real-world spatial and non-spatial datasets show that LENS consistently outperformed the state-of-the-art semi-supervised fake review detection methods at all different labeling rates on real and fake reviews, respectively.

Funding This work was supported by Sacramento State Research and Creative Activity Faculty Awards Program.

Data availability All the data/codes in the experiments are available on request.

References

1. Aghakhani, H., Machiry, A., Nilizadeh, S., Kruegel, C., Vigna, G.: Detecting deceptive reviews using generative adversarial networks. In: 2018 IEEE Security and Privacy Workshops (SPW), pp. 89–95. IEEE (2018)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning, pp. 214–223. PMLR (2017)
3. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003). URL http://jmlr.org/papers/v3/bengio03a.html
4. Croce, D., Castellucci, G., Basili, R.: GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2114–2119. Association for Computational Linguistics, Online (2020). URL https://www.aclweb.org/anthology/2020.acl-main.191
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018). URL http://arxiv.org/abs/1810.04805
6. Dieng, A.B., Ruiz, F.J., Blei, D.M.: Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics **8**, 439–453 (2020)
7. Ehsani, K., Mottaghi, R., Farhadi, A.: Segan: Segmenting and generating the invisible. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6144–6153 (2018)
8. Elman, J.L.: Finding structure in time. Cognitive science **14**(2), 179–211 (1990)
9. Fedus, W., Goodfellow, I., Dai, A.M.: Maskgan: better text generation via filling in the_. arXiv preprint arXiv:1801.07736 (2018)
10. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 171–175 (2012)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
12. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Advances in neural information processing systems **30** (2017)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Huszár, F.: How (not) to train your generative model: Scheduled sampling, likelihood, adversary? arXiv preprint arXiv:1511.05101 (2015)
15. Jindal, N., Liu, B., Street, S.: opinion-spam-and-analysis-wsdm-08. pdf. In: Proc. First ACM Int. Conf. Web Search Data Min (2008)
16. Konda, V., Tsitsiklis, J.: Actor-critic algorithms. Advances in neural information processing systems **12** (1999)
17. Kumar, A., Sattigeri, P., Fletcher, T.: Semi-supervised learning with gans: Manifold invariance with improved inference. Advances in neural information processing systems **30** (2017)
18. Li, F.H., Huang, M., Yang, Y., Zhu, X.: Learning to identify review spam. In: Twenty-second international joint conference on artificial intelligence (2011)
19. Li, H., Chen, Z., Mukherjee, A., Liu, B., Shao, J.: Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 9, pp. 634–637 (2015)

20. Li, H., Liu, B., Mukherjee, A., Shao, J.: Spotting fake reviews using positive-unlabeled learning. Computación y Sistemas **18**(3), 467–475 (2014)
21. Metz, L., Poole, B., Pfau, D., Sohl-Dickstein, J.: Unrolled generative adversarial networks. arXiv preprint arXiv:1611.02163 (2016)
22. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: What yelp fake review filter might be doing? In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 7 (2013)
23. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. CoRR **abs/1610.09585** (2016). URL http://arxiv.org/abs/1610.09585
24. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557 (2011)
25. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
26. Rayana, S., Akoglu, L.: Collective opinion spam detection: Bridging review networks and metadata. In: Proceeding of the 21st ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'15 (2015)
27. Rayana, S., Akoglu, L.: Collective opinion spam detection using active inference. In: S.C. Venkatasubramanian, W.M. Jr. (eds.) Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016, pp. 630–638. SIAM (2016). DOI 10.1137/1.9781611974348.71. URL https://doi.org/10.1137/1.9781611974348.71
28. Socher, R., Huang, E., Pennin, J., Manning, C.D., Ng, A.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. Advances in neural information processing systems **24** (2011)
29. Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D.: Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the 2011 conference on empirical methods in natural language processing, pp. 151–161 (2011)
30. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 1631–1642 (2013)
31. Stanton, G., Irissappane, A.A.: Gans for semi-supervised opinion spam detection. arXiv preprint arXiv:1903.08289 (2019)
32. Tsai, A.C., Lin, S., Fu, L.: Contrast-enhanced semi-supervised text classification with few labels. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, pp. 11394–11402. AAAI Press (2022)
33. Tuan, Y.L., Lee, H.Y.: Improving conditional sequence generative adversarial networks by stepwise evaluation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **27**(4), 788–798 (2019)
34. Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., Matsumoto, Y.: Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. arXiv preprint arXiv:1812.06280 (2018)
35. Yu, L., Zhang, W., Wang, J., Yu, Y.: Seqgan: Sequence generative adversarial nets with policy gradient. In: Proceedings of the AAAI conference on artificial intelligence, vol. 31 (2017)
36. Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820 (2015)