Fake News Detection with Generated Comments for News Articles

Yuta Yanagi
Department of Informatics
University of Electro-Communications
Tokyo, Japan
yanagi.yuta@ohsuga.lab.uec.ac.jp

Ryohei Orihara
Department of Informatics
University of Electro-Communications
Tokyo, Japan
orihara@acm.org

Yuichi Sei
Department of Informatics
University of Electro-Communications
Tokyo, Japan
seiuny@uec.ac.jp

Yasuyuki Tahara
Department of Informatics
University of Electro-Communications
Tokyo, Japan
tahara@uec.ac.jp

Akihiko Ohsuga
Department of Informatics
University of Electro-Communications
Tokyo, Japan
ohsuga@uec.ac.jp

***Abstract*—Recently, fake news is shared via social networks and makes wrong rumors more diffusible. This problem is serious because the wrong rumor sometimes make social damage by deceived people. Fact-checking is a solution to measure the credibility of news articles. However the process usually takes a long time and it is hard to make it before their diffusion. Automatic detection of fake news is a popular researching topic. It is confirmed that considering not only articles but also social contexts(i.e. likes, retweets, replies, comments) supports to spot fake news correctly. However, the social contexts are naturally unavailable when an article comes out, making early fake news detection by means of the social context useless. We propose a fake news detector with the ability to generate fake social contexts, aiming to detect fake news in the early stage of its diffusion where few social contexts are available. The fake context generation is based on a fake news generator model. This model is trained to generate comments using a dataset which consists of news articles and their social contexts. In addition, we also trained a classify model. This used news articles, real-posted comments, and generated comments. To measure our detector' s effectiveness, we examined the performance of the generated comments for articles with real comments and generated ones by the classifying model. As a result, we conclude that considering a generated comment help detect more fake news than considering real comments only. It suggests that our proposed detector will be effective to spot fake news on social networks.**

***Index Terms*—fake news, disinformation, neural network, natural language processing, deep-learning, microblogs**

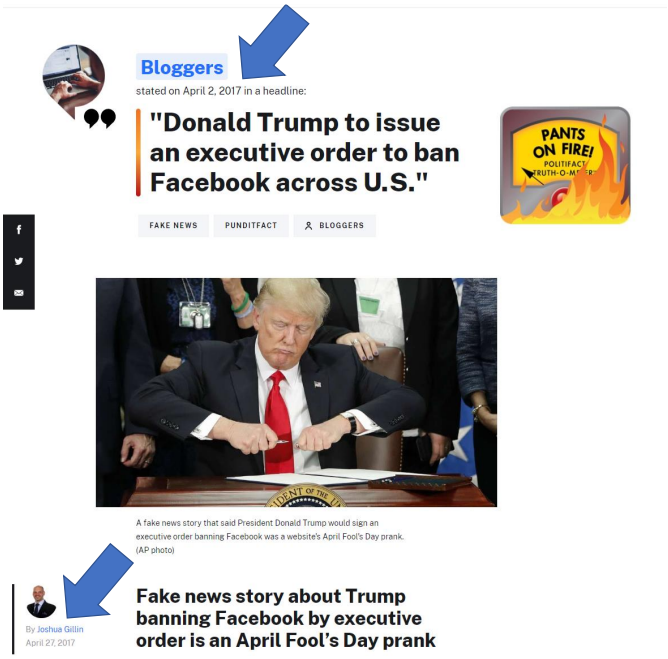

Fig. 1. An example of fact-checking. It is confirmed as an April Fool's prank. The blue arrows show the posted date of the fake news and the fact-checking result. It takes 25 days for the news to be verified as *fake*.

I. Introduction

In this era, social media is one of the important parts of our lives. Social media makes it easier to get news and share them with friends online. However, there is also information with less credibility. Some of them have misinformation that is made by malicious purposes. We call them "fake news".

Fake news tries to make false rumors diffusible by being shared. This year, there is so much fake news on COVID-19 and sometimes make wrong rumors in the social networks. Directer-General of the WHO called this problem "infodemic" and he told that fake news is shared faster and more easily than the virus itself [**?**]. Besides, fake news created some not only online, but also offline (real incidents) e.g. in Washington D.C., fake news on the Pizzagate conspiracy is reported to have motivated the shooting [**?**]. Nowadays, fact-checking is the most used method to spot fake news. This is a process of evaluating news by people who have knowledge of news topic. Fig. **??** is an example of fact-checking [**?**]. However, this takes so long that it is hard to spot a piece of fake news before it is shared. Fake news also shakes the premise of democracy due to people cannot get accurate information. Therefore, researches

try to detect fake news through machine learning.

The challenge of this work is there are news articles which try to deceive readers on purpose and this makes it harder to classify them by a simple rule-based method. Trying to supplement information for detection, there are some works which aggregate social context(i.e. retweet, like, and comments) report better results than considering news text only [**?**]. However, social contexts are unavailable before being shared. Hence, there is also a work that generates words likely to be appeared in comments from the news by CVAE to detect fake news when they are just posted [**?**]. Although their work tries to generate comments, they have only achieved to produce words that have a high probability of appearing in the comments.

In this work, we will propose a generation model that evaluates news credibility by news text and generated comments. This model is modified from generating fake news articles [**?**] and this model learns not only news features but also how to generate comments. In training, this sequence includes real posted comments but the test sequence does not use them. The skill of generating comments help classification in the test data.

We measure the performance of our proposed method by experiments with a real posted dataset.

II. Related works

To detect and classify fake news is not a new topic because it is similar to detecting spam [**?**], rumor [**?**], and false advertisement [**?**]. Following previous works [**?**], [**?**], [**?**], we define fake news as news that is intentionally fabricated and can be verified as wrong.

A. Detecting fake news

There are many works which detect fake news based on the news content only. In-text features, writing styles [**?**] and amount of emotions [**?**] were considered as promising features because commonly fake news has unique styles and emotions. Besides, using deep neural networks achieved better results in classification on previous works [**?**], [**?**], [**?**].

Many works consider the social context of news content. The social context features are generated by user-based [**?**], [**?**], [**?**], post-based [**?**], [**?**], [**?**], and network-based [**?**], [**?**].

Considering the social context, the detection must wait for a while from posting of an article because social contexts are made by users who are exposed to the article. Therefore, the Two-Level Convolutional Neural Network with User Response Generator(TCNN-URG) was proposed [**?**]. This generates comment by hidden variables which are trained by a probability distribution of comments appearance. Generated comments can give additional information to classify posts and the model is available even if the news is just posted. However, the TCNN-URG generates only words that have a high probability of appearing in a comment and it generates no grammatical elements.

B. Generating fake news

In generating natural language articles, the Grover model made natural neural fake news articles [**?**]. This model is trained by a news dataset where news articles are organized in fields such as news domain, author, posted date, title, and article. The model was evaluated by the performance of the prediction of one of the news elements. An interesting finding made by them is that human beings are more likely to be fooled by generated articles than by real ones. We tried to extend this model and generate natural comments.

III. methodology

As we saw in **??**, the original Grover model was trained by a news dataset which had five parts. Each part is attached start and end tokens. After the training, data without the tags are given to evaluate prediction performance. We replaced the fields other than the article with three comments and tried to predict one of the comments from the other fields. We modeled the generation model by the joint distribution like the original one:

$$p = (\mathrm{article}, \mathrm{comment\_1}, \mathrm{comment\_2}, \mathrm{comment\_3}) \quad (1)$$

This model's diagram is shown in Fig. **??**. Basically it was constructed by replacing fields in Grover model' s news structure with comments, except for article. The purpose of our model was to generate not articles but comments likely to be written by humans.

The last token of integrated news fields was `[CLS]` and this was used for classification into real/fake. This is a same method as for GPT-2 [**?**]. The original one was made for the generation of fake news but our proposed model was arranged to generate comments. Fig.**??** shows our process of experiment.

IV. results

A. Word generation tendency

First of all, we investigated the difference between generated comments from real and fake news. We generated comments which refer to news articles that are fact-checked by PolitiFact from the FakeNewsNet dataset [**?**]. This dataset contains sets of a news article and tweets(comments) which refer to it. We chose news articles which have at least three tweets and sampled three tweets for comment generation. We prepared 200 sets of news articles and comments for each of the real and fake classes and trained the model to generate comments. We used the following indexes to evaluate words appeared in the generated comments: the number of the occurrence of the words(shown in percentage), ratio of the word's occurrence among the total number of words, and the difference between the percentages for real and fake classes. We converted all alphabets to the lowercase letter. We removed the following elements: stop words provided by NLTK [**?**], url(starts with *http*), and symbols such as quotation, period, comma, and so on in order to investigate the frequency of not symbols but words accurately. On the other hand, we spared mentions,

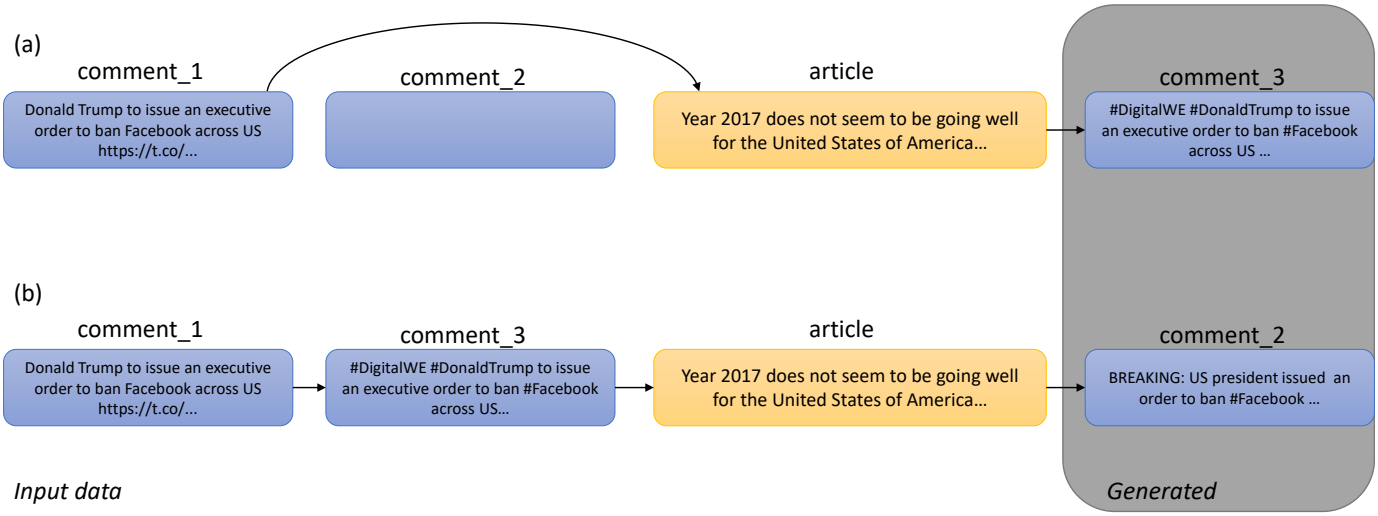

Fig. 2. Two cases of comment generation with our proposed model. (a) shows a case where a comment was generated from an article and a real-posted comment. (b) shows another case where a comment was generated from data that includes the generated comment in (a).

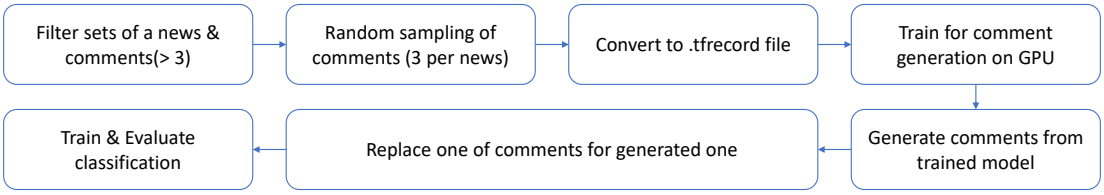

Fig. 3. The process of our experiment.

colons, and hashtags(i.e. @anyone, analyze:, #anything) because the addition of these symbols serves different purposes in the social networks. We found the following by the analysis of the generated comments:

- The most frequent word was "via"(approx. 1.5%) in the aggregated frequency in real and fake.
- The second and third were "trump" and "obama" however both of their percentages were under 1%.

We also found the following from the difference between generated comments from real and fake news.

- The word "via" was also the top frequency of generated word from both real and fake.
- The percent of the frequency of "via" in generated comments from fake news article was twice as much as ones from real news.
- The word "via" was also the word with the largest margin of the percentages between real and fake classes. The margin was approx. 0.9 point.
- The word "breaking:" was the word with the second-largest margin of the percentages between real and fake classes, where one for the fake was more than the real. The margin was approx. 0.7 point.

B. *Quality of classification*

We measured the effect of generated comments for classification by comparing classification results with and without the generated comments. We prepared baselines: classification with the news article only, and with the article and two real-posted comments. In this experiment, we used pairs of an article of GossipCop, another data available in the FakeNewsNet dataset, and tweets referring it instead of PolitiFact because the latter has too few data to make the experiment meaningful. We sampled the same rule as **??** although we collect 2000 sets each for the real and fake classes. The result of classification is Table **??**. Our proposed method achieved the best recall score however in precision it was outperformed by models that disregard the generated comments. On the whole, the generated comments seemed to be not accurate in grammar.

V. DISCUSSION

A. *Generating comments*

According to trends of words in generated comments, our proposed method seemed to be trained by the topics of news articles. Most of the generated comments referred to topics of politics and this may be caused by the character of the dataset.

TABLE I
RESULTS OF CLASSIFICATION

Model name	Precision	Recall	F1 score
Article only	0.647	0.615	0.631
+ Real comment * 2	**0.682**	0.750	**0.714**
+ Generated comment	0.590	**0.790**	0.675

An interesting word in the generated comments is "breaking:". Our experimental results showed that the word was generated more by fake news than real news. The phenomenon was not reported in the research of TCNN-URG [**?**]. Their research claimed that "!", "?", "false", and so on were important signals of fake news. "Breaking:" maybe also signal of fake news.

The grammatical quality of the generated comments was clearly poor. This is caused by a lack of dataset scale. Grover model was built using 120 gigabytes of dataset [**?**]. We need to search or get a larger dataset of articles and tweets.

B. classification

According to TABLE **??**, our proposed model achieved the best recall score however in precision its performance was worst. This means the proposed model can detect more fake news than another model which disregards generated comments even if available social contexts are limited.

The trend suggests that this model helps people who search for news which require fact-checking. However, the model also detected more false fake news than another one therefore we need to make an improvement. We will check if the trend is changed by using a larger dataset.

APPENDIX
SETTINGS OF EXPERIMENTS

- Trained on Ubuntu 16.04 on Docker in Linux server with TITAN X (Pascal).
- Our proposed model was extended from Grover repository by forking on GitHub.
- Model size was Grover-Base but we reduced vocabulary a little bit in order to fit for the extension.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Numbers JP17H04705, JP18H03229, JP18H03340, JP18K19835, JP19K12107, JP19H04113.