Reproducibility Report: Sentimental features in fake news data set

Yixuan Wu
Eberhard Karl University of Tübingen
yixuan.wu@student.uni-tuebingen.de

Abstract

The reproducibility crisis (also known as replication crisis) is a crucial issue existing in the scientific methodology, which describes the fact that many of the empirical results from scientific studies failed to be replicated or reproduced. As an engineering side of computational linguistics, the reproducibility of natural language processing (NLP) has attracted more interests. Fake claims detection is a task which mostly uses the NLP techniques, and as people constantly receive various of information in their daily life, fraud news detection also has become an important approach to filter out the untruthful news pieces. This paper aims at testing the reproducibility of a paper from Upadhayay and Behzadan (2020) which brought up an advanced fraud news dataset and a model for fake news detection, and discussing about the factors which may cause the variation in the result.

1 Introduction

For many years, scientific studies mainly involved convincing details about how researches are conducted and carried out, so other researchers can replicate the result. A real scientific result should be replicable, because replications can help to verify the reliability of science, allow more people to build on the scientific results, and eventually help to move a field developing further - science which cannot be replicated eventually becomes obsolete.

However, in practice, most of the published studies rarely achieved replication, which is understandable as it is more exciting to produce new results than repeating the old ones. For field like physics, chemistry or computer science, replication is usually expected since the researchers (if not all) have provided the information needed to rerun the experiment. Though the replications are rare, more and more researchers start to realise that there exist quite a few reported findings which could not be reproduced (Collaboration, 2015; Freedman et al., 2015). In fact, 70% of the scientists have reported that they failed to reproduce the published findings from other people, and more than 50% have reported difficulty or failure to even reproduce their own work (Baker, 2016). Such a phenomenon is known as the "reproducibility crisis". In 2019, Mieskes et al. (2019) reported that there were 24.9% of attempts to reproduce the team's original work, and 56.7% of attempts to reproduce the work from other scientists have been reported as failure. In the area of natural language processing (NLP), more voices start to express the concern of the reproducibility over the past years (Mieskes, 2017; Reimers and Gurevych., 2017; Belz et al., 2021).

In recent years, a growing number of workshops and conferences in machine learning (ML) and NLP fields have focused on addressing the challenges and importance of reproducibility, and reproducibility seems to gradually become a standard part in the reviewing process (Belz et al., 2021). Nowadays, in ML/NLP communities and most of the main conferences have started to encourage and make space for sharing the code, data, and all the necessary supplementary materials, to ensure the validity of the results (Belz et al., 2021).

In the paper from Upadhayay and Behzadan (2020), a task of fake news detection was addressed. In the paper, Upadhayay and Behzadan (2020) brought up a more advanced fake news data set - SentimentLIAR (the old version is the original LIAR data set) and a BERT-Based model which eventually reached around 70% accuracy on detecting the fake news from the SentimentLIAR data set. This paper will try to replicate the work from Upadhayay and Behzadan (2020), then conduct another fraud news detection task by using the LIAR data set. The results will be compared together, and

the encountered problems will be discussed in the following sections.

2 Scope of reproducibility

This paper aims at reproducing the result from a fake news detection work conducted by Upadhayay and Behzadan (2020), who claimed to have built an advanced counterfeit news detection dataset and a model for detecting fraud information. Fake news detection task is the classification of a given text to test if the embedded information is misinformation or not. The output is either 0 or 1, with the fake news being marked as 0 and real news marked as 1.

3 Methodology

The reproduction process of this paper consists of the following steps: (1) using the same code which is shared by Upadhayay and Behzadan (2020) to reproduce reported result (minor changes were added in the script due to the update of used packages); (2) applying the old version of dataset to the same model proposed by Upadhayay and Behzadan (2020) and compare the result with the result from step 1. For the sake of distinguishing the two main steps contained in this study, the term "reproduction" is used to refer to exactly reproducing the work from Upadhayay and Behzadan (2020); the term "replication" is adopted to refer to applying the original LIAR data set in fake news detection.

3.1 Dataset

The data set and the corresponded script for the replication work is obtained from the repository provided by Upadhayay and Behzadan (2020) [1]. This SentimentalLIAR data set is the modified and further extended version of the original LIAR data set [2] which was proposed by Wang (2017). In the LIAR data set, there are 14 columns contained: column 1: the ID of the statement, column 2: label, column 3: statement, column 4: subject(s), column 5: the speaker, column 6: the speaker's job title, column 7: the state info, column 8: the party affiliation, column 9: barely true counts, column 10: false counts, column 11: half true counts, column 12: mostly true counts, column 13: pants on fire counts, and column 14: the context (venue / location of the speech or statement). In a word, column 9-13 contain the information of the total credit history count, including the current statement. An example is provided in Table 1.

ID	153.json
Label	half-true
Statement	"I'm the only person on this stage who has worked actively just last year passing, along with Russ Feingold, some of the toughest ethics reform since Watergate."
Subject	ethics
Speaker	Barack Obama
Job Title	President
State Information	Illinois
Party affiliation	Democrat
Barely-true counts	70
False counts	71
Half true counts	160
Mostly true counts	163
Pants on fire counts	9
Context	a Democratic debate in Philadelphia, Pa.

Table 1: Example of an item in LIAR.

There are two main differences between LIAR and SentimentLIAR: (1) in the LIAR data set, there are 6 different labels to determine the degree of truthfulness of a text - "True", "mostly true", "half true", "barely true", "false", and "pants fire"; whereas in the SentimentLIAR data set, all the above 6 categories are reduced to 2 levels, with "half-true", "false", "barely-true" and "pants-fire" labels were marked as "False", and all the rest labels were marked as "True" (such change is represented in a binary style, with [0, 1] represents False and [1, 0] represents True). For the replication work, such a binary feature was also added into the LIAR data set;(2) in the SentimentLIAR dataset, more features/columns were added - Upadhayay and Behzadan (2020) utilized the Google NLP API [3] to assign the sentiment "positive" or "negative" labels to the statement according to the numeric score (these change is tagged as the "SEN"

[1] https://github.com/UNHSAILLab/SentimentalLIAR
[2] https://github.com/tfs4/liar_dataset
[3] https://cloud.google.com/natural-language

feature for its sentiment representation); besides, Upadhayay and Behzadan (2020) also added 5 different emotion states to each text: anger, sadness, disgust, fear and joy ("EMO" feature). The score of each emotion state is calculated through IBM NLP API [4]. In Upadhayay and Behzadan (2020)'s work, the speaker's name is converted into a numerical ID in order to prevent the bias of textual name representations (Upadhayay and Behzadan, 2020). Statement, subject, speaker ID, job title, state information, party affiliation and sentiment label were concatenated as the TEXT attribute, the credits of the speaker were chained as SPC attribute, and the scores in different emotional fields were integrated as the EMO feature. Besides of these changes, the rest parts remain identical to the LIAR data set. An example is provided in Table 2.

3.2 Model Description

As claimed by Upadhayay and Behzadan (2020), the best performance they got was from the BERT-Base + CNN model, with the TEXT fed into the BERT-Base, and the EMO, SPC and SEN integrated together with the output from BERT-Base, and then fed jointly into the CNN. Such a proposed model was adopted in this study as well.

Consider the differences between LIAR and SentimentLIAR (information in SEN and EMO attributes were newly added thus only exist in SentimentLIAR), when conducting the replication work by using LIAR, the TEXT attribute was concatenated with everything identical to the TEXT in SentimentLIAR except the speaker_id was replaced by the speaker information (e.g., "_3_" in SentimentLIAR TEXT, "Barack Obama" in LIAR TEXT). The information in SPC feature remains the same in both the reproduction and replication works, and the EMO and SEN attributes were not included in the LIAR based replication work.

3.3 Hyperparameters

The hyperparameters used in this study were almost identical to the ones recorded in the published paper by Upadhayay and Behzadan (2020), except that in the LIAR based replication work, the max_length was shrinked to 250 whereas in the SentimentLIAR reproduction work, the max_length was 300. Such change is made to fit the hardware of the compiling environment for the replication work. This change could affect the number of the truncated instances being fed into the model.

[4]https://www.ibm.com/cloud/watson-natural-language-understanding

TEXT	Statement	"I'm the only person on this stage who has worked actively just last year passing, along with Russ Feingold, some of the toughest ethics reform since Watergate."
	Subject	ethics
	speaker_id	_2_
	Job Title	President
	State Information	Illinois
	Party affiliation	Democrat
	Sentiment	NEGATIVE
SPC	Barely-true counts	70
	False counts	71
	Half true counts	160
	Mostly true counts	163
	Pants on fire counts	9
EMO	anger	0.021023
	fear	0.077569
	joy	0.032182
	disgust	0.038037
	sad	0.438594
SEN	sentiment score	-0.20000000298023224

Table 2: Example of an item in SentimentLIAR.

4 Results

In the first part of the experiment, the result from Upadhayay and Behzadan (2020) was attempted to be replicated and compared with the published result. In fact, by rerun the original script, the result is slightly different from the published one.

According to the report from Upadhayay and Behzadan (2020), the best performed BERT-Base

+ CNN model eventually reached around 70% accuracy and 0.6430 F1 score. The replication result, however, showed a slightly decrease in both scores. The difference can be seen in Table 3.

	Published result	**Replication result**
Accuracy	0.6992	0.6789
F1 score (macro)	0.6430	0.6204

Table 3: Published result and the replicated result.

Considering that the random number of generating the input sequence can affect the final outcome, the original script and the replication script both applied the same number in random_state (random_state = 200) when creating the training, test, and validation data sets. All the other parameters were also remained the same as the ones reported by Upadhayay and Behzadan (2020). Though the replicated results showed that both the accuracy and F1 score were not identical to the reported ones, it is expected that the eventual replication result should be identical to the published one.

In the second part of the experiment, an attempt was conducted to get the fraud news detection result by using the original LIAR data set, this step is necessary in order to test if the SentimentLIAR data set is really more practical and advanced than the LIAR data set. Similar to the method from Upadhayay and Behzadan (2020), at first the TEXT information was fed into the Bert-Base, then instead of concatenationg the EMO, SPC and SEN information together, only the SPC information was combined with the output from BERT-Base, and then passed into the CNN (this is because EMO and SEN information was lacked in the LIAR data set but exist in the SentimentLIAR data set). Another change is that the parameter max_length was reduced from 300 to 250 in the experiment with LIAR data set. For a fair comparison between the two data sets, the script provided by Upadhayay and Behzadan (2020) was rerun with the max_length also reduced to 250, then the accuracy and F1 score in both attempts were compared. The result of comparison is presented in Table 4.

	LIAR	**Sentiment LIAR**
Accuracy	0.6945	0.6828
F1 score (macro)	0.5847	0.6431

Table 4: Comparison between LIAR and SentimentLIAR after reducing max_length to 250

5 Discussion

The experiments above has accomplished the aim of this study. For the reproduction work of Upadhayay and Behzadan (2020), the result was not the same as the published one, even when the parameters and the applied data set were controlled to be the same, with the training, test, and validation set were divided in the same ratio, and the random status in both Upadhayay and Behzadan (2020) and the reproduction work were all remained untouched. It is not clear what caused the difference, since Upadhayay and Behzadan (2020) did not report the platform they applied - the system architecture, operating system, Python version and libraries' versions were not clear. The difference could be caused by any factor listed above, but this study did not investigate further on the reason behind it.

For the replication work based on LIAR data set, the same parameters were applied except the max_length was reduced from 300 to 250. With the max_length in the SentimentLIAR script also reduced to 250, the result indicates that by using SentimentLIAR, a higher F1 macro score was observed (0.6431 when using SentimentLIAR, 0.5847 when using LIAR), though the accuracy was slightly lower than using the LIAR data set (0.6828 when use SentimentLIAR, 0.6945 when use LIAR). Considering that number of items in the classes of "True" and "False" are imbalanced in both the LIAR and SentimentLIAR datasets (with around 65% "False" items and 35% "True" items), the comparison was more focused on the F1 macro score instead of the accuracy. The result suggests that with the comparison of F1 score, SentimentLIAR does provide a better performance than the original LIAR data set in detecting fake news.

In sum, the reproduction procedure was not so complicated thanks to the detailed script provided by Upadhayay and Behzadan (2020), only a few changes were conducted in this study due to the incompatibility in package versions. Though the reproduced result by using SentimentLIAR is slightly different from the published result, this study confirms the fact that by using SentimentLIAR and the proposed BERT-Base + CNN architecture instead of LIAR, a better performance in distinguishing fake news can be achieved. Such an outcome seems

to point out that sentiment score is a potential feature which plays a role in distinguishing real or fabricated information.

References

Monya Baker. 2016. Reproducibility crisis. *Nature*, 533(26):353–66.

A. Belz, S. Agarwal, A. Shimorina, and E. Reiter. 2021. A systematic review of reproducibility research in natural language processing. arXiv preprint arXiv:2103.07929.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 1(349):943–951.

Leonard P. Freedman, Iain M. Cockburn, and Timothy S. Simcoe. 2015. The economics of reproducibility in preclinical research. *PLOS Biology*, 13(6):1–9. 06.

Margot Mieskes. 2017. A quantitative study of data in the nlp community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, page 23–29, Valencia, Spain. Association for Computational Linguistics.

Margot Mieskes, Karëen Fort, Cyril Grouin Auréelie Névéol, and Kevin Cohen. 2019. Community perspective on replicability in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, page 768–775, Varna, Bulgaria. INCOMA Ltd.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 338–348, Copenhagen, Denmark, September. Association for Computational Linguistics.

Bibek Upadhayay and Vahid Behzadan. 2020. Sentimental liar: Extended corpus and deep learning models for fake claim classification. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.

W. Y. Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection.