Replication of the Reliability Prediction for Health-Related Content: A Replicability Study

TOBIAS ABRAHAM HAIDER, Vienna University of Technology, Austria
WANJIA TANG, Vienna University of Technology, Austria
MIRJANA SADIKOVIKJ, Vienna University of Technology, Austria
ANUM HAFEEZ, Vienna University of Technology, Austria

This study aims to replicate the work of Sondhi and his colleagues, who proposed a method for determining the reliability of online health-related content using support vector machines based on a set of linguistic and web-specific features extracted from web sites. Our replication study made use of the code provided by the authors and aimed to make it more reproducible by updating outdated libraries and replacing custom implementations with standard ones. We were able to obtain results that were consistent with the original study, and our replication effort has resulted in a refactored version of the original code that is now also available to the research community. Our replication effort confirmed the conclusions of the original study and highlighted some challenges involved in replication of research.

Additional Key Words and Phrases: support vector machine, natural language processing, reliability prediction, health, machine learning, reproducibility

ACM Reference Format:
Tobias Abraham Haider, Wanjia Tang, Mirjana Sadikovikj, and Anum Hafeez. 2024. Replication of the Reliability Prediction for Health-Related Content: A Replicability Study. 1, 1 (December 2024), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 THE REPRODUCED PAPER

The paper "Reliability Prediction for Health-Related Content: A Replicability Study" [1] presents a method for determining the reliability of online health-related content using standard classification technology and a set of linguistic and web metadata features. The main goal of the original paper is to build a document-level classifier using a supervised learning approach to detect false information on the internet.

1.1 Methods used in the reproduced paper

1.1.1 Dataset used in the reproduced paper. The authors manually created a fully balanced dataset with reliable and unreliable web pages that was used in the replicability task. The dataset consisted of positive pages from websites accredited by HON (Health On The Net) and negative pages from a web search using hand-crafted queries.

Authors' addresses: Tobias Abraham Haider, e11833743@student.tuwien.ac.at, Vienna University of Technology, Karlsplatz 13, Vienna, Vienna, Austria, 1040; Wanjia Tang, e123456789@student.tuwien.ac.at, Vienna University of Technology, Karlsplatz 13, Vienna, Vienna, Austria, 1040; Mirjana Sadikovikj, e123456789@student.tuwien.ac.at, Vienna University of Technology, Karlsplatz 13, Vienna, Vienna, Austria, 1040; Anum Hafeez, e11833743@student.tuwien.ac.at, Vienna University of Technology, Karlsplatz 13, Vienna, Vienna, Austria, 1040.

XXXX-XXXX/2024/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.1.2 Features generated in the reproduced paper. The features used include link-based features (number and type of links, presence or absence of privacy policy information or contact links), commercial features (presence of commercial interest and advertisements) and linguistic features (type-token ratio, average sentence length, etc).

1.1.3 Classification approach chosen in the reproduced paper. In the original paper, support vector machines are used to build a classifier to detect false information on the web.

1.2 Original results from the reproduced paper

2 PROJECT SETUP

The original paper provided a link to a GitHub repository containing the code used to run the experiments. The code was all in one file called train.py and was designed to be run using the command "python3 train.py" with multiple parameters for the training, such as the dataset, features, and standardization. A requirements.txt and readme file were also provided, which made it appear easy to run the experiment as is.

3 CHALLENGES

Running the code on different machines presented several challenges. One issue was the use of the name "aux" as a directory for storing intermediary training data. This keyword is forbidden in Windows, making it impossible to run the code on a Windows machine.

Another challenge was the dependencies in the code itself, which needed to be installed and run before the experiment. These included libraries that needed to be installed using pip commands, a library called svmlight that needed to be installed by downloading and manually extracting it on Linux, and tokenizers for punctuations and stop words that needed to be downloaded for the library nltk.

4 EXPERIMENT CHANGES

4.1 Changes made to File paths

The name of the directory used for storing intermediary training data was changed from "aux" to "train_data" to make the code runnable on Windows machines.

4.2 Changed libraries and implementations

The code was refactored into one Jupyter notebook. This way, the rather complex project setup could be performed and documented in one place. This included the installation of various libraries using pip, the download of the tokenizers.

The most notable change compared to the original code is the used implementation of the support vector classifier. It was switched to the more modern implementation found in sklearn and could therefore be used without a separate manual installation. With the change of the svc dependency, a custom implementation of the confusion matrix function was removed, since it is already present in the sklearn package.

5 REPRODUCED RESULTS

In Fig 1, We have gotten the same results as the one (Table 6) from the paper corresponding with the same SVM cost factor values. Unfortunately, one of our limitations with reproducing the result is in processing the results for the CLEF eHealth dataset, which requires high storage. As we do not have access to a supercomputer, we have not produced the results for CLEF Health dataset. In Fig 2, We implemented the same technique for the Sondhi dataset and got similar significant feature results as the ones from Table 1.

Fig. 1. Schwarz

					weighted accuracy(%).		
Features	SVM cost factor	F1	F1 (reliable class)	F1 (non reliable class)	1	2	3
Links	1	**0.9375**	**0.97**	**0**	**93.75**	—	—
	2	0.9375	0.97	0	—	88.26	—
	3	0.9375	0.97	0	—	—	88.26
Links + Commercial	1	**0.9375**	**0.97**	**0**	**93.75**	—	—
	2	0.9375	0.97	0	—	88.26	—
	3	0.9375	0.97	0	—	—	83.4
Words (removing stopwords)	1	0.9125	0.95	0	91.25	—	—
	2	0.9125	0.95	0	—	85.88	—
	3	**0.9125**	**0.96**	**0.25**	**—**	**—**	**84.39**
Words (keeping stopwords)	1	0.9375	0.97	0	93.75	—	—
	2	0.9125	0.95	0	—	85.88	—
	3	0.9125	0.95	0	—	—	81.13
All (removing stopwords)	1	**0.9375**	**0.97**	**0**	**93.75**	—	—
	2	0.9125	0.95	0	—	85.88	—
	3	0.9125	0.95	0	—	—	81.13
All (keeping stopwords)	1	0.9125	0.95	0	91.25	—	—
	2	**0.9375**	**0.97**	**0.45**	—	**90.59**	—
	3	**0.9375**	**0.97**	**0.45**	—	—	**87.8**

Fig. 2. Sondhi

					weighted accuracy(%)		
Features	SVM cost factor	F1	F1 (reliable class)	F1 (non reliable class)	1	2	3
Links	1	0.74	0.69	0.77	73.61	—	—
	2	0.74	0.67	0.78	—	80.19	—
	3	**0.74**	**0.66**	**0.78**	**—**	**—**	**84.31**
Links + Commercial	1	0.78	0.77	0.79	78.06	—	—
	2	0.72	0.62	0.78	—	80.74	—
	3	**0.72**	**0.61**	**0.78**	**—**	**—**	**86.11**
Words (removing stopwords)	1	0.64	0.44	0.73	63.61	—	—
	2	0.65	0.44	0.74	—	76.67	—
	3	**0.65**	**0.44**	**0.74**	**—**	**—**	**82.5**
Words (keeping stopwords)	1	0.59	0.3	0.71	58.61	—	—
	2	0.59	0.31	0.71	—	72.78	—
	3	**0.59**	**0.31**	**0.71**	**—**	**—**	**79.58**
All (removing stopwords)	1	0.64	0.44	0.73	63.61	—	—
	2	0.65	0.46	0.74	—	76.67	—
	3	**0.65**	**0.46**	**0.74**	**—**	**—**	**82.5**
All (keeping stopwords)	1	0.59	0.3	0.71	58.61	—	—
	2	0.59	0.31	0.71	—	72.78	—
	3	**0.59**	**0.31**	**0.71**	**—**	**—**	**79.58**

6 INTERPRETATION OF THE RESULTS

Due to machine capabilities, we could test only for the Sondhi and Schwarz datasets.

Fig 1 presents the experiment's results evaluating the reliability of different approaches and different datasets with different features/parameters.

It can be seen that the Schwarz dataset using all features (allKeep and allRem) had the highest accuracy, precision, recall, and F1 score, particularly in the second and third iterations. The dataset using only words (wordsKeep and wordsRem) had the second-highest accuracy, precision, recall, and F1 score, particularly in the first iteration. The approach using only links (link) had the lowest accuracy, precision, recall, and F1 score. It appears that the Schwarz dataset using all available features (allKeep and allRem) is the most promising approach for distinguishing reliable from unreliable sites, as it achieved the highest performance across all iterations. Sondhi datasets has similar resulting significant features too.

7 SUMMARY

In conclusion, the current study replicates the original research on reliable technology and confirms the effectiveness of word-based models and those that incorporate multiple features in distinguishing reliable from unreliable websites. Further testing on diverse datasets supports the validity of these findings. Additionally, the algorithm's ability to successfully address the issue of COVID-19 misinformation in the TREC 2020 Health Misinformation Track further demonstrates its generalizability. The research community can access the code for replication at Github.

REFERENCES

[1] Marcos Fernández-Pichel, David E. Losada, Juan C. Pichel, and David Elsweiler. 2021. Reliability Prediction for Health-Related Content: A Replicability Study. In *Advances in Information Retrieval*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer International Publishing, Cham, 47–61.