

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

TOBIAS ABRAHAM HAIDER, WANJIA TANG , MIRJANA SADIKOVIKJ, ANUM HAFEEZ
and 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 Our replication study made use of the code provided by the authors and aimed to make it more reproducible by updating outdated libraries and replacing custom implementations with standard ones.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, Anum Hafeez, Tobias Abraham Haider, Wanjia Tang, Mirjana Sadikovikj, and Anum Hafeez. 2023. Replication of the Reliability Prediction for Health-Related Content: A Replicability Study. 1, 1 (January 2023), 2 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 THE REPRODUCED PAPER

The paper "Reliability Prediction for Health-Related Content: A Replicability Study" 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

Authors' addresses: Tobias Abraham Haider, Wanjia Tang , Mirjana Sadikovikj, Anum Hafeez; 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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.
XXXX-XXXX/2023/1-ART \$15.00
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

websites accredited by HON (Health On The Net) and negative pages from a web search using hand-crafted queries.

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

3 CHALLENGES

4 EXPERIMENT CHANGES

4.1 Changes made to File paths

4.2 Changed libraries and implementations

5 REPRODUCED RESULTS

6 INTERPRETATION OF THE RESULTS

7 SUMMARY

REFERENCES