FACULTY OF DESIGN, MEDIA AND INFORMATION
Department of Information

RAT
RESULT ASSESSMENT TOOL

HAW
HAMBURG

Web Page Classification: A Systematic Literature Review

A Master's Thesis by Nurce Yagci
Faculty of Design, Media and Information, Department of Information

Background

With the rapid growth and sharing of information on the internet, the need to automatically identify the topic or sentiment of a web page or social media post has become necessary to meet the demands of systems and users alike (Ullah et al., 2020). Most research focuses on classifying web pages or documents in one domain, like phishing detection or identification of fake news. These studies focus on the content of web pages. Another aspect that helps describe a document, however, is genre. Orlikowski and Yates define a genre as "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form" (Orlikowski & Yates, 1994, p. 543). This definition is refined by describing genre as sharing a common form, content, or purpose (Crowston, 1997). By knowing the genre of the document, a person can be more precise when formulating their queries and more appropriately judge the relevance of presented results. Similarly, systems also benefit from knowing the genre of a document when evaluating its relevance in association with a query (Andersen, 2008).

Research Question

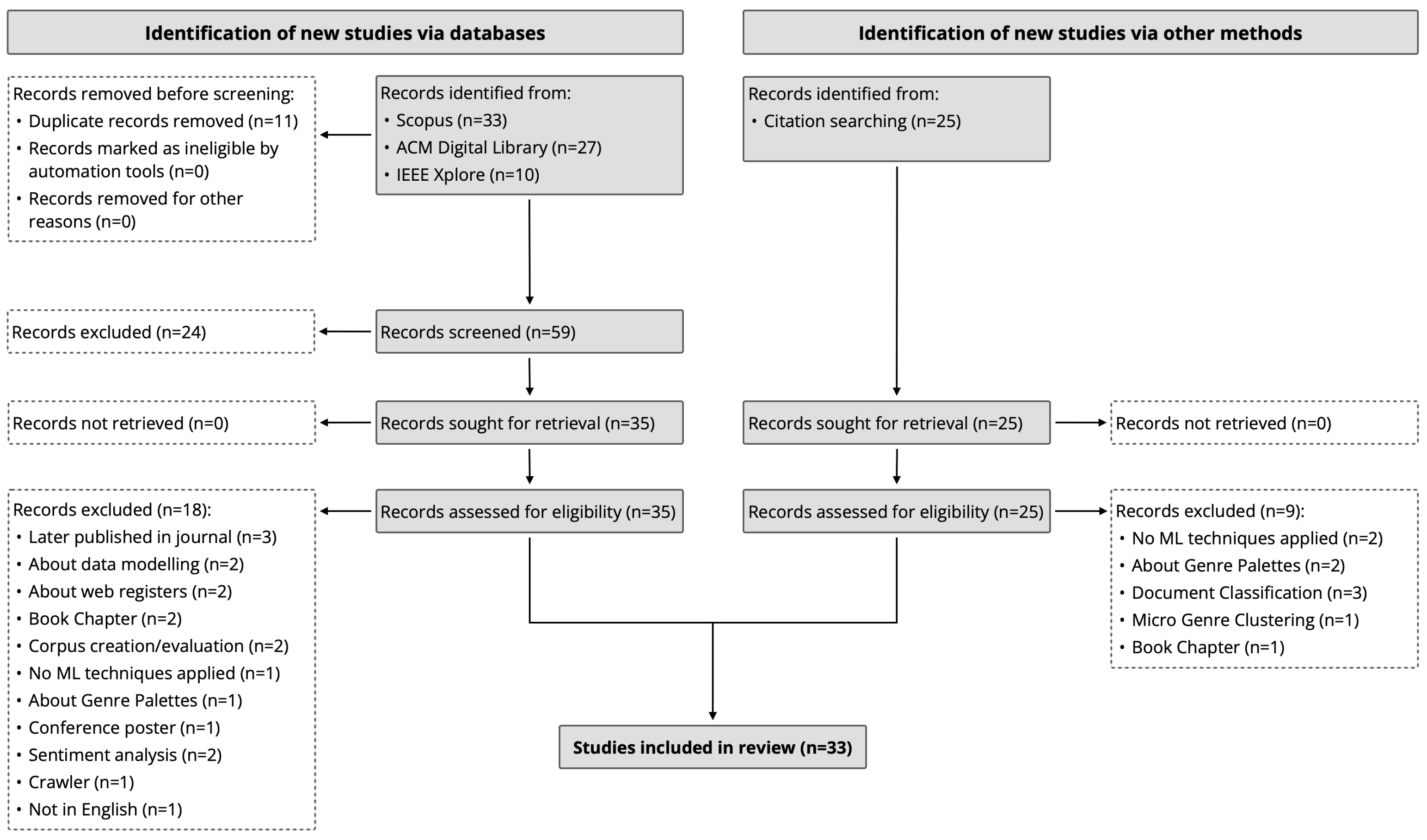
What is the state-of-the-art research on web genre classification using machine learning techniques?

Each step in the process of common machine learning approaches is used to formulate subquestions regarding the (1) data sets, (2) output features, (3) input features, (4) pre-processing methods, (5) classification algorithms, (6) hyper-parameter tuning methods, (7) validation strategies, (8) evaluation metrics, (9) overall performance, (10) strengths and weaknesses of each study.

Method

The PRISMA reporting protocol describes several steps that need to be taken to conduct a good SLR. The first step requires the formulation of inclusion and exclusion criteria, which help to identify relevant studies. The next step is the selection of information sources that will be used to find the studies. The databases ACM Digital Library, IEEE Xplore and Scopus were selected (Tieppo et al., 2022; Wen et al., 2012). Additionally, following the identification of relevant studies from the databases, both backward and forward snowballing was performed (Wohlin, 2014).

The search query was formulated using significant terms from the research question and including synonyms and allowing alternative spellings. The general query was `("web* genre*") AND (classif* OR detect* OR identif*) AND ("machine learning" OR autom*)` and has been adapted to fit the syntax of each database. Studies were selected by applying the eligibility criteria. Next, information from the selected studies was extracted. In the last step, ten questions were used to evaluate the quality of the selected studies.


Identification of new studies via databases

Identification of new studies via other methods

Records removed before screening:
- Duplicate records removed (n=11)
- Records marked as ineligible by automation tools (n=0)
- Records removed for other reasons (n=0)

Records identified from:
- Scopus (n=33)
- ACM Digital Library (n=27)
- IEEE Xplore (n=10)

Records identified from:
- Citation searching (n=25)

Records excluded (n=24)

Records screened (n=59)

Records not retrieved (n=0)

Records sought for retrieval (n=35)

Records sought for retrieval (n=25)

Records not retrieved (n=0)

Records excluded (n=18):
- Later published in journal (n=3)
- About data modelling (n=2)
- About web registers (n=2)
- Book Chapter (n=2)
- Corpus creation/evaluation (n=2)
- No ML techniques applied (n=1)
- About Genre Palettes (n=1)
- Conference poster (n=1)
- Sentiment analysis (n=2)
- Crawler (n=1)
- Not in English (n=1)

Records assessed for eligibility (n=35)

Records assessed for eligibility (n=25)

Records excluded (n=9):
- No ML techniques applied (n=2)
- About Genre Palettes (n=2)
- Document Classification (n=3)
- Micro Genre Clustering (n=1)
- Book Chapter (n=1)

Studies included in review (n=33)

Figure 1. PRISMA 2020 flow diagram (Page et al., 2021)

Results

The most common datasets used to train AWGC models are the KI-04 and 7-Genre corpora. These datasets contain eight and seven web genres, and 1,209 and 1,400 web pages, respectively. The most common input sources are web pages, text and HTML tags. NLP methods (n-grams, part-of- speech tagging) and frequency-based features (number of links and average sentence length) are used for feature generation. Most approaches implement minimal pre-processing strategies (stopword removal).

Generally, approaches use fewer features, but in the case of more features, selection methods (Information Gain, TF-IDF) are applied to select the most relevant features. The most commonly used algorithm is Support Vector Machines (SVM). Most studies do not adjust the hyper-parameters of the algorithms, however when applied Grid Search is implemented.

Ten-fold cross-validation, Accuracy, and F-measure are used to validate and measure the performance of AWGC models. The performances for single-label classification models are excellent, but the mean performance of more sophisticated approaches is lower (86% Accuracy, 50% F-measure).

Discussion

The need for more and updated sample data is evident. Most of the corpora in use have been annotated by up to three people and contain less than 1,500 samples. A large corpus would be a more realistic training set.

Furthermore, this corpus should include noise and hierarchical relationships and allow pages to have multiple genres. Current benchmark data sets only contain a low number of genres that have not been updated for almost twenty years.

References:
1. Ullah, M. A., Tahrin, A., & Marjan, S. (2020). An Algorithm for Multi-Domain Website Classification. International Journal of Web-Based Learning and Teaching Technologies, 15(4), 57– 65. https://doi.org/10.4018/IJWLTT.2020100104
2. Orlikowski, W. J., & Yates, J. (1994). Genre Repertoire: The Structuring of Communicative Practices in Organizations. Administrative Science Quarterly, 39(4), 541. https://doi.org/ 10.2307/2393771
3. Crowston, K. (1997). Reproduced and emergent genres of communication on the World-Wide Web. Proceedings of the Thirtieth Hawaii International Conference on System Sciences, 6, 30–39. https:// doi.org/10.1109/HICSS.1997.665482
4. Andersen, J. (2008). The concept of genre in information studies. Annual Review of Information Science and Technology, 42(1), 339–367. https://doi.org/10.1002/aris.2008.1440420115
5. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, n71. https://doi.org/10.1136/bmj.n71
6. Tieppo, E., Santos, R. R. dos, Barddal, J. P., & Nievola, J. C. (2022). Hierarchical classification of data streams: A systematic literature review. Artificial Intelligence Review, 55(4), 3243–3282. https://doi.org/10.1007/s10462-021-10087-z
7. Wen, J., Li, S., Lin, Z., Hu, Y., & Huang, C. (2012). Systematic literature review of machine learning based software development effort estimation models. Information and Software Technology, 54(1), 41–59. https://doi.org/10.1016/j.infsof.2011.09.002
8. Wohlin, C. (2014). Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. https://doi.org/10.1145/2601248.2601268

Contact
nurce.yagci@haw-hamburg.de

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences