ISSI 2023 Proceedings, v2, pp 181–187
DOI: 10.5281/zenodo.8370893 https://cns-iu.github.io/workshops/2023-07-02_issi/

Appropriateness of citing retracted articles in biomedicine: Sentiments expressed in citations without acknowledgement of retraction

Tzu-Kun Hsiao[1]

[1]*tkhsiao2@illinois.edu*
University of Illinois at Urbana-Champaign, School of Information Sciences, 501 E. Daniel St, Champaign, Illinois, 61820, (USA)

Abstract

Citations to retracted articles after they have been retracted (i.e., post-retraction citations) can be problematic if the retracted articles are cited as legitimate work. To gain a deeper understanding of how problematic post-retraction citations are, we analyzed the sentiments expressed in 3,156 post-retraction citation contexts to see whether the retracted articles were cited positively as legitimate work after their retractions. Our results showed that the vast majority of post-retraction citations cited retracted articles as legitimate work: 84.27% (2,660 out of 3,156) of the post-retraction citation contexts lacked acknowledgement of retraction and expressed positive sentiments. We also investigated the potential to automatically detect the sentiment. To evaluate how well sentiment could be automatically detected, supervised machine learning models based on logistic regression, support vector machine (SVM), convolutional neural network (CNN), and bidirectional long short-term memory (BiLSTM) were developed. The best-performing model was a CNN model augmented with sentence embeddings and hand-crafted features (0.79 accuracy and 0.60 macro F1). Our findings indicate that detecting citation sentiment is a challenging task. The improvement obtained from augmenting the word embeddings model with other features shows that sentence embeddings and hand-crafted features extracted from text similarity and a sentiment lexicon capture additional sentiment cues.

Introduction

Although retraction is commonly regarded as a sign of problematic research, empirical studies have shown that retraction does not prevent the retracted article from spreading (Dinh et al., 2019; Szilagyi et al., 2022). Citations to retracted articles may persist for a long time after their retractions (Szilagyi et al., 2022). In studies on citations to retracted articles, post-retraction citations are frequently regarded as problematic because citations imply using the retracted work in the citing work. However, not all citations are positive. For example, citations can be used to critique or refute the cited work (Xu et al., 2015). Therefore, whether a post-retraction citation is truly problematic should be assessed by how the retracted article is cited in the full text. This includes (1) the reason for citing retracted work, (2) citation sentiment, and (3) whether the authors acknowledged the retraction. Some studies have investigated the sentiments expressed in citations to retracted articles, but the scales of analysis was limited. Bar-Ilan and Halevi (2017) studied 238 citations to 15 retracted articles and reported most citations were positive. Theis-Mahon & Bakker (2020) analyzed 685 citations to 81 retracted dentistry articles and reported that 69.3% of the citations were positive. Yang et al. (2022) examined citations to 98 retracted psychology articles and found that 90.84% were positive.

In a previous study (Hsiao & Schneider, 2021), we systematically examined acknowledgments of retraction using 13,252 post-retraction citation contexts (i.e., sentences containing citations) identified from 2 million PubMed Central (PMC) open-access articles (Hsiao & Torvik, manuscript in preparation). Only 5.4% (722) of the 13,252 post-retraction citation contexts acknowledged the retractions. In order to gain a deeper understanding of how problematic post-retraction citations are, this study builds on our prior research by (1) investigating sentiments expressed in post-retraction citation contexts lacking acknowledgments and (2) exploring the possibility of automatically identifying the sentiments. This article reports our ongoing work on annotating post-retraction citation sentiments and designing a model for automatic sentiment identification. The scale of the annotations and the database-wide coverage of citation contexts contribute to a more comprehensive picture of post-retraction citation sentiment. In our experiments for the current article, we used features based on different text representations, text similarity, and a sentiment lexicon to build the model. The results of the experiments provide insight into detecting sentiments from various perspectives.

Methodology

Data collection and sentiment annotations

Using the PubMed IDs (PMIDs) of the retracted articles, 13,252 post-retraction citation contexts referencing 2,763 retracted articles were identified from a dataset containing 2,049,871 PMC open-access articles (Hsiao & Schneider, 2021). 12,530 (94.55%) of the 13,252 citation contexts did not acknowledge the retractions (Hsiao & Schneider, 2021). These 12,530 citation contexts are the subject of this study. To date, we (TKH) have annotated sentiments for 3,156 citation contexts (25.19% of the total 12,530 post-retraction citations without acknowledgment of the retraction), using 4 categories as shown in Table 1: s*trongly positive, weakly positive*, *neutral*, and *negative*.

Table 1. Citation sentiments

Sentimen	Definition	Example
Strongly positive	Citing work uses something (e.g., definitions, concepts, materials, equipment, and techniques) in the retracted work. Citing work confirms, is supported by, depends on, or explicitly agrees with the retracted work.	Similar to our results, SLA mRNA expression levels were previously reported to be downregulated in human B-cells and were strongly expressed in naïve, pre-germanal center, and germinal-center B-cells based on gene expression analysis. (Retracted PMID: 12438421).
Weakly positive	Retracted work is cited as related and legitimate, without raising concerns.	Altered NPM1 expression was observed in many types of tumors, and mutated NPM1 is frequently detected in human hematopoietic malignancies, especially in acute myeloid leukemia (AML) (Retracted PMID: 18401421).
Negative	Citing work disputes, corrects, questions, or disagrees with the retracted work. Citing work expresses concerns or casts doubts on the retracted work. Citing work indicates that the retracted work has shortcomings or uncertainty. Citing work mentions that the findings reported in the retracted work are controversial or lacking confirmation.	The original report of an alphaproteobacterial Sphingomonas-related GAO (Retracted PMID: 15256569) was later shown to be incorrect and the FISH probes were shown to be binding to members of the Defluviicoccus cluster 1.
Neutral	Retracted work is cited without additional comment. No judgment of validity is shown in the citation context.	The contribution of inflammatory cytokines to tumor development has been investigated by other studies (Retracted PMID: 25544369, Non-retracted PMIDs: 18954521, 18036640)

Features for designing an automatic sentiment identification model

The following features were extracted and tested to develop a model for automatic sentiment identification.

Bag-of-Words (BOW) Features. Stanza, a natural language processing (NLP) toolkit released by the Stanford NLP group (Qi et al., 2020), was used to extract words in citation contexts, as well as their lemmas (the canonical form of words) and part-of-speech (POS) tags. Lemmas were used as features to handle different expressions of a word (e.g., *report*, *reports*, *reported*, and *reporting*). In preliminary experiments, BOW features were constructed using all the lemmas, but models built with this feature set performed poorly. One of the possible explanations is that sentiments are often hinted at in verbs, adjectives, and adverbs. In the example below, for instance, the sentiment was revealed mainly by the adjective, "*incorrect*".

An example of a negative citation context (from PMID: 28496434):

> "The original report of an alphaproteobacterial Sphingomonas-related GAO (Retraction PMID: 15256569) was later shown to be incorrect and the FISH probes were shown to be binding to members of the Defluviicoccus cluster 1."

Furthermore, nouns in citation contexts were often topic-related terms that were unrelated or weakly related to sentiments. To address this issue, the BOW features used in this study were filtered using POS tags, which provided grammatical context for words. For example, the word "*report*" may have a POS tag as NOUN (e.g., a lab *report*) or VERB (e.g., the authors *report* their findings). Our filtering procedure consisted of two steps: (1) We excluded lemmas tagged as proper nouns, numbers, punctuations, and symbols (2) For lemmas tagged as nouns, we retained only a selected set of nouns (see Appendix) in the BOW features. These nouns were chosen because their co-occurrence with the citing/cited party or sentiment was frequently observed during the annotation process. The retained lemmas were concatenated with their POS tags to capture the word sense ambiguity to some extent. That is, each feature in the BOW features was a lemma-POS pair (e.g., report VERB).

Word and Sentence Embeddings. BOW is indifferent to word order and is not ideal for handling phrases. For instance, the meaning of the phrase "*Down syndrome"* may be lost since this phrase is represented in BOW features as two independent words, "*down*" and "*syndrome*". To address the limitations of BOW features, word and sentence embeddings were generated to serve as contextual text representations because they could better capture the implicit semantics expressed in words and sentences (Chen et al., 2019; Zhang et al., 2019). BioWordVec (Zhang et al., 2019) and BioSentVec (Chen et al., 2019) were used to generate these embeddings, because they are publicly available word and sentence embeddings pre-trained on similar data: articles in PubMed and clinical notes in the MIMIC-III Clinical Database.

Sentiment Scores. Kilicoglu et al.'s (2019) lexicon was used to calculate the sentiment score for each citation context. Unigrams, bigrams, and trigrams in each citation context were mapped to positive and negative terms in the lexicon. For each of the mapped terms, the sentiment score of the term in the citation context was the reciprocal of the distance (D) between the mapped term and the citation marker pointing to the retracted article (i.e., 1/D for a positive term and -1/D for a negative term). The distance was defined as the number of words between the mapped term and the citation marker. The purpose of this design was to reflect word order based on the assumption that the closer the term was to the citation marker, the more likely the sentiment inferred from the term was related to the retracted article. A citation context's sentiment score was the sum of the sentiment scores of the mapped terms.

Pairwise Cosine Similarity. Pairwise cosine similarity between citation contexts was calculated using the BioSentVec sentence embeddings of each pair of contexts. The assumption was that citation contexts having the same sentiment would have a higher cosine similarity.

Machine-learning models tested for automatic sentiment identification

Experiments using traditional machine learning and deep learning models were performed to explore how well post-retraction citation sentiments could be automatically identified. Model performance was evaluated using 10-fold cross-validation.

Traditional machine-learning models. Experiments were conducted using logistic regression and support vector machine (SVM) since these two models had promising results in predicting sentiment (Kilicoglu et al., 2019; Xu et al., 2015). To evaluate the predictive power of the feature sets, the model was first trained on BOW features only, and the remaining feature sets were added to the model one by one.

Deep learning models. Convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) were also chosen since they had shown promising results in text classification tasks (Kilicoglu et al., 2019; Kim, 2014). We began the experiments with word embeddings as model inputs because they are the most common feature in studies that use CNN and BiLSTM models for citation sentiment classification (Yousif et al., 2019). Although word embeddings could capture some contextual information, they did not capture all the cues pointing to sentiments (Kilicoglu et al., 2019). To test whether additional features could improve model performance, we experimented with augmenting the models with sentence embeddings, pairwise cosine similarity, and sentiment scores, under the assumption that these additional features could provide informative cues that word embeddings could not capture.

Results

Sentiments expressed in the post-retraction citation contexts

Of the 3,156 annotated citation contexts, 84.28% (2,660) were positive (2,078 weakly positive and 582 strongly positive), and 6.56% (207) were neutral. Only 289 (9.16%) citation contexts were negative. Recall that the annotated citation contexts are those not acknowledging the retraction. The dominance of positive sentiments highlights the problem that retracted articles are frequently cited as legitimate work without informing the readers about the retractions.

Current progress on automatic sentiment identification

We discovered that the deep learning models CNN and BiLSTM outperformed the traditional machine learning models. Among the traditional machine learning model, the best performance was from a logistic regression model with BOW, sentence embeddings, sentiment scores, and pairwise cosine similarity as features. This logistic regression model had an accuracy of 0.62 and a macro F1 score of 0.53, while the best deep learning model had an accuracy of 0.79 and a macro F1 score of 0.60. The findings indicate that deep learning models captured more nuance in citation contexts with different sentiments. The improvements in the augmented models (Table 2) demonstrate that the additional features did provide useful sentiment cues that word embeddings could not capture.

Table 2. Accuracy and Macro F1 of the Deep Learning Models

Model	Features	CNN		BiLSTM	
		Accuracy	F1	Accuracy	F1
Base model	Word embeddings	0.77	0.51	0.78	0.53
Augmented model 1	Word embeddings + Sentence embeddings	**0.79**	**0.60**	0.78	0.59
Augmented model 2	Word embeddings + Sentence embeddings + Pairwise cosine similarity + Sentiment scores	**0.79**	**0.60**	0.78	0.59

Table 3 shows the performance of the best deep learning models (i.e., the two augmented CNN models with 0.79 accuracy and 0.60 macro F1) by sentiment categories. Although the best models had the same overall performance, there were differences in their ability to identify each sentiment. The CNN with sentence embeddings, pairwise cosine similarity, and sentiment scores slightly outperformed the CNN with only sentence embeddings. According to the precision, recall, and F1 scores, this CNN model performed better in detecting *negative*, *neutral*, and *strongly positive* sentiments. These results indicate that the additional inputs (pairwise cosine similarity and sentiment scores) captured sentiment cues that word and sentence embeddings could not. This improvement also implies that citation contexts with the same sentiment may be more similar to each other in their text representations and may have a tendency to have terms in the lexicon with the same sentiment direction.

Table 3. Best Model Performance Per Category

Model	Features	Overall		Per category			
		Accuracy	Macro F1	Sentiment	Precision	Recall	F1
Augmented CNN model 1	Word embeddings + Sentence embeddings	0.79	0.60	Negative	0.59	0.31	0.39
				Neutral	0.32	0.06	0.10
				Weakly positive	0.80	**0.94**	0.86
				Strongly positive	0.73	0.65	0.68
Augmented CNN model 2	Word embeddings + Sentence embeddings + Pairwise cosine similarity	0.79	0.60	Negative	**0.62**	**0.53**	**0.57**
				Neutral	**0.38**	**0.22**	**0.26**
				Weakly positive	**0.84**	0.91	**0.88**
				Strongly	**0.74**	**0.67**	**0.70**

Discussion and future works

This study investigated how problematic post-retraction citations were by analyzing the sentiments expressed in post-retraction citation contexts that did not acknowledge the retraction. Compared to prior work focusing on a particular field (Theis-Mahon & Bakker, 2020; Yang et al., 2022) or a small number of retracted articles (e.g., 15 retracted articles in Bar-Ilan and Halevi (2017)), this work broadened the scope of understanding of post-retraction citation sentiment through a database-wide analysis of citation contexts. Our results align with previous studies (Bar-Ilan & Halevi, 2017; Yang et al., 2022), finding that citations to retracted articles were mostly positive: 84.28% (2,660/3,156) of the analyzed post-retraction citation contexts were positive. This raises the concern that most post-retraction citations could be problematic since retracted articles were most often cited as valid work, without mentioning the retraction.

We conducted experiments to develop a machine learning model for automatically detecting sentiments expressed in citation contexts. This model could benefit future studies analyzing how retracted articles were cited in full-text articles in the biomedical field. The current state of progress demonstrates that identifying sentiments from citation contexts is a challenging task. Although challenges such as handling grammar errors, irregular spellings, and slang are less common in academic writing, the avoidance of criticizing others' work in citations (MacRoberts & MacRoberts, 1984; Xu et al., 2015) make sentiment cues less obvious. The 0.79 accuracy and 0.60 macro F1 achieved by the best CNN model indicate room for further improvement. Nonetheless, the experiments provide insights for capturing the subtle sentiment cues expressed in text: (1) Sentence embeddings provided informative perspectives that were different from word embeddings, and (2) the improvement achieved by augmenting the CNN model with pairwise cosine similarity and sentiment scores implied that citation contexts with the same sentiment might have shared characteristics (e.g., using negations (*not* adequate, *not*... as previously indicated) to express a negative sentiment).

This study makes two main contributions: First, this is the first database-wide study on post-retraction citation sentiments. Second, the improved experiment results show that cues derived from features other than word embeddings help sentiment detection. For future work, our priority is to finish annotating the remaining 9,374 post-retraction citation contexts. The completed annotations will further shed light on how retracted articles are cited in the biomedical literature. Second, we will continue the experiments on building the model by incorporating other features (e.g., dependency-enhanced word embeddings in Kilicoglu et al. (2019)) and by testing CNN model variants (Kim, 2014) that may further improve the model performance.

Acknowledgements and author contributions

This work was partially funded by Alfred P. Sloan Foundation G-2020-12623. CRediT Roles: Conceptualization–TKH. Data curation–TKH and Jodi Schneider. Formal analysis–TKH. Methodology–TKH. Writing–original draft–TKH. Writing–review & editing–TKH and Jodi Schneider.

References

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: A case study. *Scientometrics*, *113*(1), 547–565. https://doi.org/10.1007/s11192-017-2242-0

Chen, Q., Peng, Y., & Lu, Z. (2019). BioSentVec: Creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics*, 1–5. https://doi.org/10.1109/ICHI.2019.8904728

Dinh, L., Sarol, J., Cheng, Y.-Y., Hsiao, T.-K., Parulian, N., & Schneider, J. (2019). Systematic examination of pre-and post-retraction citations. *Proceedings of the Association for Information Science and Technology*, *56*(1), 390–394. https://doi.org/10.1002/pra2.35

Hsiao, T.-K., & Schneider, J. (2021). Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies*, *2*(4), 1144–1169. https://doi.org/10.1162/qss_a_00155

Hsiao, T.-K., & Torvik, V. I. (manuscript in preparation). OpCitance: Citation contexts identified from the PubMed Central open access articles.

Kilicoglu, H., Peng, Z., Tafreshi, S., Tran, T., Rosemblat, G., & Schneider, J. (2019). Confirm or refute?: A comparative study on citation sentiment classification in clinical research publications. *Journal of Biomedical Informatics*, *91*, 103123. https://doi.org/10.1016/j.jbi.2019.103123

Kim, Y. (2014). Convolutional Neural Networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

MacRoberts, M. H., & MacRoberts, B. R. (1984). The negational reference: Or the art of dissembling. *Social Studies of Science*, *14*(1), 91–94. https://doi.org/10.1177/030631284014001006

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108.

Szilagyi, I.-S., Schittek, G. A., Klivinyi, C., Simonis, H., Ulrich, T., & Bornemann-Cimenti, H. (2022). Citation of retracted research: A case-controlled, ten-year follow-up scientometric analysis of Scott S. Reuben's malpractice. *Scientometrics*, *127*(5), 2611–2620.

Theis-Mahon, N. R., & Bakker, C. J. (2020). The continued citation of retracted publications in dentistry. *Journal of the Medical Library Association*, *108*(3), 389–397.

Xu, J., Zhang, Y., Wu, Y., Wang, J., Dong, X., & Xu, H. (2015). Citation sentiment analysis in clinical trial papers. *AMIA Annual Symposium Proceedings*, *2015*, 1334–1341.

Yang, S., Qi, F., Diao, H., & Ajiferukea, I. (2022). Do retraction practices work effectively? Evidence from citations of psychological retracted articles. *Journal of Information Science*, Advance online publication. https://doi.org/10.1177/01655515221097623

Yousif, A., Niu, Z., Chambua, J., & Khan, Z. Y. (2019). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, *335*, 195–205. https://doi.org/10.1016/j.neucom.2019.01.021

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, *6*(1), 52.

Appendix

Nouns included in the BOW features: accordance, addition, agreement, analysis, author, concern, contrast, controversy, correlation, example, experiment, evidence, failure, finding, hypothesis, improvement, instance, limitation, literature, method, model, other, report, research, result, study, suspicion, work.