Cite-worthiness Detection on Social Media: a Preliminary Study

Salim Hafid[1[0000−0002−1775−8542]], Wassim Ammar[1], Sandra Bringay[1,2[0000−0002−2830−3666]], and Konstantin Todorov[1[0000−0002−9116−6692]]

[1] LIRMM, CNRS, University of Montpellier, France
[2] University Paul Valéry, France
{firstname.lastname}@lirmm.fr

Abstract. Detecting cite-worthiness in text is seen as the problem of flagging a missing reference to a scientific result (an article or a dataset) that should come to support a claim formulated in the text. Previous work has taken interest in this problem in the context of scientific literature, motivated by the need to allow for reference recommendation for researchers and flag missing citations in scientific work. In this preliminary study, we extend this idea towards the context of social media. As scientific claims are often made to support various arguments in societal debates on the Web, it is crucial to flag non-referenced or unsupported claims that relate to science, as this promises to contribute to improving the quality of the debates online. We experiment with baseline models, initially tested on scientific literature, by applying them on the SciTweets dataset which gathers science-related claims from X. We show that models trained on scientific papers struggle to detect cite-worthy text from X, we discuss implications of such results and argue for the necessity to train models on social media corpora for satisfactory flagging of missing references on social media. We make our data publicly available to encourage further research on cite-worthiness detection on social media.

Keywords: Cite-worthiness · Science-related discourse · Social Media · NLP.

1 Introduction

Social media, especially X (ex-Twitter), has become a vital platform for scientific discourse among scholars, but also among non-academic users. Scientists rely on X as a convenient platform for sharing findings and connecting with peers [30], while non-scientific users often call upon scientific results or formulate science-related claims in order to give more weight to their arguments in societal debates on various, often controversial topics. For example, discussions surrounding the recent COVID-19 global pandemics were often fueled by science-related arguments—verified or not—relating to vaccines efficiency or protection measures. While a lot of attention has been given to analysing science-related

claims from scientific literature [23], only recently the natural language processing (NLP) community started taking interest in scientific discourse on social media and on the Web at large [22]. These recent efforts have been largely motivated by the observation that scientific discourse is arguably different on social media as compared to academic literature, where social media users leaning on science in their discourse would often lack rigour, oversimplify or mis-contextualize scientific findings [21].

A specific problem in that context is that of cite-worthiness detection, seen as the task of *"identifying citing sentences, i.e., sentences which contain a reference to an external source"* in text [1]. In particular, this task can be useful for flagging a missing reference to a scientific result (an article or a dataset) that should come to support a claim formulated in the text, hence giving credit to the original author, giving credibility to the claim presented or providing additional insights. Previous work has taken interest in this problem in the specific context of scientific literature, motivated by the need to allow for reference recommendation for researchers and to flag missing citations in scientific work [1]. In our work, we extend this idea towards the context of social media, leveraging the results and models reported in [1]. While scientific claims are often made to support various arguments in societal debates on the social Web, the lack of citation standards, as compared to academic writing, leads to the presence of largely unsupported science-related claims and mis-contextualized scientific findings, which in turn leads to a poor quality of the debates online, lacking transparency, credibility and accuracy, with potentially harmful effects on democratic discourse [15–17].

In [1], several well-known pre-trained language models, such as SciBERT and Longformers are fine-tuned for the specific task of cite-worthiness detection in scientific literature and evaluated against a simple logistic regression baseline, by relying on data tailored for the task.[3] In our preliminary study, we follow the protocol provided by [1], by applying and fine-tuning the same models and baselines, but in contrast using data coming from X exclusively. Namely, we rely on the SciTweets dataset [3],[4] which gathers human annotated science-related claims from X, based on the definition of scientific web claims and the annotation protocol given in [3]. We further preprocess and filter tweets from SciTweets to map them to the cite-worthiness definition from [1]. We observe consistent decline across all metrics when evaluating models on X data. This hints that the inherent difference between academic and social media scientific discourse [13, 14, 18] translates to a degraded performance of baseline models on the downstream task of cite-worthiness detection, calling for specific models that are capable of taking into consideration the specificity of scientific discourse on the Web.

In this work, we contribute:

[3] https://github.com/copenlu/cite-worth
[4] https://github.com/AI-4-Sci/SciTweets

1. SCiteTweets, the first publicly available dataset for cite-worthiness detection on social media, consisting in 415 tweets constructed by preprocessing and filtering tweets from the SciTweets dataset.[5]
2. The first empirical evaluation of cite-worthiness detection on social media, where we observe that performance of models trained on scientific publications consistently declines when evaluated on data from X.

2 Related Work

The notion of cite-worthiness relates to the notion of check-worthiness, which has been extensively researched by fact-checking related studies over the years.[6] A sentence is defined as "check-worthy" if it is worth fact-checking (e.g., contains a verifiable factual claim, is potentially harmful, and is of general interest) [29, 28], whereas a sentence is "cite-worthy" if it contains a reference to an external source [1]. While check-worthiness detection can help professional fact-checkers detect which claims to focus on, cite-worthiness detection can be used to flag scientific results which are presented without references.

Determining whether a (scientific) text lacks and hence requires a citation, has been one of the challenges in the NLP community. The larger group of approaches has tackled this problem in the context of scientific publishing, using corpora constructed from academic articles in specific fields. For example, [24] use Support Vector Machines on a dataset created from the ACL Anthology Reference corpus [25], while more advanced approaches [6] measure the performance of a Convolutional Recurrent Neural Network on the ACL Arc dataset[7] as well as arXivCS [26] and Scholary Dataset.[8] The limitations of these works are mainly related to domain-specificity, class imbalances, and little to no presence of data quality analysis. These issues were addressed in [1], where the authors build and share a curated multi-domain dataset specifically dedicated to the task of cite-worthiness detection, that is used to evaluate a number of language models against a logistic regression baseline.

On the social media side, existing work [7] observed that the nature of X has led to a more lenient way of citing, especially in the scientific field where discourse is expected to be more formal. The larger amount of work analysing X data and scientific discourse is generally about the lack of trust in the shared content [8], more precisely focusing on fact-checking. For example, in [9] the authors create a manually annotated dataset to identify claims as check-worthy, while in [10] the authors leverage Large Language Models to build datasets for identifying misinformation.

Studying scientific citation in social media is a relatively novel task. In [11], the authors suggest that tweets can predict the citation of papers in the biomedi-

[5] The data is made publicly available at https://github.com/SalimHFX/SCiteTweets/
[6] See the CheckThat! Lab editions hosted by the CLEF conference- https://checkthat.gitlab.io/clef2024/task1/
[7] https://paperswithcode.com/dataset/acl-arc-1
[8] https://www.db.soc.i.kyoto-u.ac.jp/ sugiyama/Dataset2.html

cal field, concluding that X citations may be an alternative to traditional ones on the impact of research findings. Supporting that work, [12] assembles a dataset relating tweets and citations of arXiv papers. Finally, [3] presents a definition of scientific web claims and provide a curated dataset of tweets annotated according to that definition. This dataset, although limited in size, provides hints about citation tendencies in scientific discourse on X.

In an attempt to provide preliminary insights into this under-researched problem, we build on the work of [1] by reproducing their experiments on X-provenance data using the SciTweets dataset from [3] in order to highlight the shortcomings of state-of-the-art pre-trained models when taken out of the academic literature context, which in turn hints to the inherent difference of discourse on social media as compared to scientific papers.

3 Data

To evaluate cite-worthiness performance on social media, we use the following two distinct datasets (examples from each dataset are shown in Table 1):

- **CiteWorth** [1]: To our best knowledge, CiteWorth is the largest dataset dedicated to cite-worthiness detection from scientific-publication text. It is extracted from the S2ORC dataset [5] which consists of 81.1M english-language scientific publications. It is then filtered, where sentences are given "cite-worthy" labels indicating that they originally contained a citation at the end of the sentence. The final dataset consists of 1.1M sentences, where over 375k sentences are labeled as cite-worthy.
- **SciTweets** [3]: SciTweets is a dataset dedicated to online scientific discourse, where authors developed a hierarchical definition of science-relatedness and curated ground-truth data from X. Tweets are categorized into different categories of science-relatedness depending on whether they contain scientific knowledge, a reference to scientific knowledge, or are related to scientific research in general. The final dataset consists of 1,261 human-annotated tweets. We use the SciTweets dataset to construct **SCiteTweets**, our dataset for cite-worthiness detection on X, by mapping SciTweets labels to cite-worthiness labels. We explain this procedure in detail in in Section 4.1.

Table 1. Samples from the existing labels in both datasets used in our experiments

	Size	Labels	Examples
CiteWorth	1,181,793	Cite-worthy	The success rate of PNA in the literature varies from 79-100%.
		Non Cite-worthy	We compared visual electrophysiology recording of patients with the normal range as defined in our laboratory.
SciTweets	1,261	Scientific knowledge	also cancer is virtually incurable bc all cancers are different :)
		Referece to scientific knowledge	Modeling precision treatment of breast cancer looks great ! http://t.co/4XzfGlwAWn
		Related to scientific knowledge	Lupus Research Institute Awards $1-Million Grants to Discover What Causes Lupus http://t.co/aXopNmLyI7
		Non science-related	These birds won't stop cherping!

4 Experiments

4.1 Setting

To evaluate the performance of existing cite-worthiness detection models on a social media corpus, we run multiple experiments to achieve the following goals: (1) reproducing the results found by authors of the CiteWorth dataset [1], (2) applying those models on the SciTweets dataset [3] to evaluate the performance of existing cite-worthiness detection models on a social media corpus, where we experiment with training models on the CiteWorth dataset and on the SciTweets datasets. To reproduce results from the CiteWorth dataset [1], we pick the following three models which all have been previously used by the authors: a logistic regression model, which represents the simplest explainable baseline, a SciBERT model [2] which had the best precision score in the authors' experiments, and a Longformer model [4] which achieved the best F1 score in the authors' experiments. While in their experiments authors used two distinct versions of Longformer, Longformer-Ctx where they use sequence modeling to embed entire paragraphs, and Longformer-Solo, where they embed single sentences, in this paper we opted to use Longformer-Solo (embedding single sentences only), as it best fits the tweets' inherently short format.

Prior to conducting the experiments, we needed to further preprocess the SciTweets dataset in order to ensure a correct mapping between its labels and the cite-worthiness labels from CiteWorth. While CiteWorth contains sentence texts and labels pointing to whether the text is cite-worthy or not, SciTweets' texts are multi-labeled. The first step was to select a label from SciTweets which can be qualified as equivalent to the cite-worthiness label from CiteWorth. The structure of the SciTweets multi-labeled dataset is as follows: a tweet is either science-related or not, if it is science-related, then the tweet is further categorized as belonging to one or more of the following subcategories: *"cat. 1: containing a scientific claim"*, *"cat. 2: containing a reference to scientific knowledge"*, or *"cat. 3: related to scientific research in general"* [3]. The first two categories are good candidates, as they can both contain cite-worthy text. However category 2 is the most suited since it references an external source of scientific nature, much like how authors constructed the CiteWorth dataset, where they focused on sentences that have an indication of a citation which is in essence an external scientific reference. Furthermore, we selected the remaining science-related tweets (categories 1 and 3) as our negative class. By doing so, we ensure that both our positive and negative classes contain science-related text, and that the classes only differ in cite-worthiness, thus matching the CiteWorth setup. The implications of this choice will be discussed further in section 5.

Moreover, we also preprocessed the tweets to match the CiteWorth setup, where we removed user-handles and URLs from cite-worthy tweets. We also removed "citation markers" at the end of sentences, as defined by authors of CiteWorth, where a citation marker is *"any text that trivially indicates a citation, such as the phrase "is shown in""*. Authors argue that removing such citation markers prevents models from learning and using these signals for prediction.

To do so, we removed excess punctuation and hanging words (*"by"*, *"via"*). This step was possible due to how the Category 2 of SciTweets [3] was constructed, where the URLs direct to actual scientific articles. We call the resulting dataset SCiteTweets, which contains tweets extracted from SciTweets that were preprocessed and filtered as described above to match the cite-worthiness definition from [1]. We show statistics of SCiteTweets in Table 2, and examples of cite-worthy and non cite-worthy sentences from both datasets in Table 3.

Table 2. Data used for the experiments

Labels	CiteWorth	SCiteTweets
Cite-worthy	375,388	207
Non Cite-worthy	806,405	208

Table 3. Examples of cite-worthy and non cite-worthy sentences from scientific papers (CiteWorth) and from tweets (SCiteTweets)

	Cite-Worthy	**Non Cite-Worthy**
CiteWorth[1]	The known forms of terrestrial life involve carbon-based chemistry in liquid water.	We compared visual electrophysiology recording of patients with the normal range as defined in our laboratory.
SCiteTweets	Hopes raised for cancer treatment after experiments halted tumour growth in mice.	proper preparation prevents poor performance

We run *three distinct experiments*, **(1)** training and evaluating on the CiteWorth dataset. This experiment is a direct reproduction of results from CiteWorth authors [1]; **(2)** training on CiteWorth and evaluating on SCiteTweets. This experiment enables us to evaluate whether models trained on a large amount of cite-worthy sentences extracted from scientific publications translates to a good performance on cite-worthy sentences from social media; **(3)** training and evaluating on SCiteTweets. This experiment enables us to evaluate whether models trained on a small amount of social media data translates to a good performance on cite-worthy sentences from social media. For each experiment, we use three distinct base-models (Logistic Regression, SciBERT, and Longformer), thus amounting to nine experiments in total. We then evaluate using Precision (P), Recall (R), and F1-score (F1) for each experiment. For all models we reproduce the experimental setting of the CiteWorth paper [1], for transformer-based models we train models on 3 epochs and follow authors' settings for all hyperparameter values such as batch size, learning rate and dropout probability. For the Logistic Regression model we use a C value of 0.11 following authors. Since the

amount of social media data we have is limited (See Table 2), we run a 10-fold cross-validation of SCiteTweets data for experiments (2) and (3). For experiment (2), the training set is the same in all folds (model is trained on CiteWorth) and only the evaluation set changes in each fold. For experiment (3), both training and evaluation sets change in each fold. We follow the same train-test split size as CiteWorth in each fold. The same seed is used for cross-validating experiments (2) and (3), thus ensuring that models are evaluated on the same test sets between the two experiments.

Table 4. Experimental results. For each model, three experiments were run, corresponding to experiments (1), (2) and (3) as described in Section 4.1

	Experiments		**Metrics**		
Models	**Trained**	**Evaluated**	**P**	**R**	**F1**
Logistic Regression	CiteWorth	CiteWorth	46.65	64.85	54.26
	CiteWorth	SCiteTweets	49.38	47.43	**48.83**
	SCiteTweets	SCiteTweets	56.11	57.83	**56.95**
SciBERT	CiteWorth	CiteWorth	65.60	52.08	58.06
	CiteWorth	SCiteTweets	53.29	23.90	**32.99**
	SCiteTweets	SCiteTweets	76.91	70.24	**73.42**
Longformer	CiteWorth	CiteWorth	56.85	68.03	61.94
	CiteWorth	SCiteTweets	54.34	23.89	**33.18**
	SCiteTweets	SCiteTweets	34.26	19.91	**25.18**

4.2 Results

We show the results of all experiments in Table 4. For experiments **(2)** and **(3)**, the presented scores are averages across 10 folds. The results of experiment **(1)** (reproducing CiteWorth results) were satisfactory, as they closely mirrored the findings outlined in the CiteWorth paper [1]. The results of experiment **(2)** (training on CiteWorth and evaluating on SCiteTweets) show a consistent decline in F1-points across all three models (LR, SciBERT, Longformer) compared to experiment **(1)**. For the baseline LR model, the decrease is of roughly 5 F1 points. For the SciBERT model, the decrease is more pronounced, with the Recall and F1 score halving compared to experiment **(1)**, recording a decrease of over 25 F1 points. And for the Longformer model, the decrease is even more noticeable, where the model loses close to 30 F1 points when evaluated on tweets compared to its performance on scientific articles.

Finally, the results of experiment **(3)** (training and evaluating on SCiteTweets) showed that most models performed best on tweets when trained on tweets. More specifically, models perform better on tweets when trained even on a small amount of tweets (experiment **(3)**), than when trained on a large amount of scientific papers (experiment **(2)**). Moreover, Longformer, the best performing model from experiment **(1)**, i.e., the best performing model on scientific papers,

is the worst performing model on tweets, having even worse scores than experiment **2**. Finally, the SciBert model outperformed both LR and Longformer on the tweets dataset.

4.3 Discussion

First, we attribute the performance of the Longformer model on experiment **(3)** to the small data size of the tweets data. We hypothesize that further experiments on a larger scale social media dataset would result in the Longformer model performing best on tweets when trained on tweets, as observed for the SciBERT and LR models (data size limitations are discussed in Section 5). Secondly, the consistent decline in F1-points across all three models (LR, SciBERT, Longformer) when training on CiteWorth and evaluating on SCiteTweets (compared to training and evaluating on CiteWorth) may be explained by differences in the linguistic structure of scientific text on the Web, where science, as discussed on the Web, differs in language from traditional scientific text from scientific papers. Existing literature has shown that scientific text on the Web uses a specialized language [13, 14], while communication studies have shown that scientific knowledge online is often sensationalized, lacks perspective [18], and has a tendency to favor conflict [19]. To verify this in our data, we show word clouds of cite-worthy text from both tweets and scientific papers in Figure 1. Cite-worthy sentences from scientific papers show a high usage of terms such as *"may"*, *"however"*, which might indicate a more careful contextualized phrasing of scientific results and of the scope in which they are valid. In contrast, cite-worthy sentences from tweets do not show usage of such terms, which might indicate a more straightforward and possibly decontextualized phrasing of scientific findings on social media. We leave for future research a more thorough analysis of linguistic differences between scientific papers text and social media text with regards to cite-worthiness.

The conclusions from the experiments in this preliminary study are that transformer-based models fine-tuned on sentences from scientific papers do not perform satisfactory on tweets for the task of cite-worthiness detection, making it difficult to correctly identify cite-worthy and check-worthy tweets, a step which has been stated by professional fact-checkers in a survey [20] as one of the main challenges and the most useful tasks to automate. In future work, we want to investigate the usefulness of training transformer-based models on larger social media corpora, with the goal of enhancing citation detection performance on social media.

5 Limitations

One limitation of our study is the size of our tweets dataset (SCiteTweets, extracted from SciTweets [3]). While the experimental results do underline a clear trend (i.e., that models trained on scientific papers underperform on scientific text from X), our results have to be cemented by further experiments on larger

Fig. 1. Word clouds for cite-worthy sentences from tweets (left) and from scientific papers (right)

scale datasets. However, to our best knowledge, the SciTweets dataset we used is the only currently existing dataset whose labels can be mapped to a cite-worthiness detection task. Another limitation is the mapping from SciTweets labels to CiteWorth labels, where we used tweets which contain a *reference to scientific knowledge* in order to match the cite-worthiness definition from the CiteWorth authors [1]. With this definition, we hope to flag tweets where users missed including references but nonetheless used language showing that there is such a reference, e.g., the following tweet with no actual reference URL: *"I read a recent study which shows that vaccines are not efficient"*. This use-case is already useful and necessary, as recent literature [27] showed that these so-called informal references are prominent on X and are shared and engaged with on social media twice as much as the actual research articles they implicitly refer to. However, ultimately, we also want to be able to flag tweets where the reference is missing and where users never meant to include it, e.g., the following tweet: *"Vaccines are not efficient"*. We consider the use-case discussed in this paper as a necessary first step towards flagging missing references on social media, and we leave the second use-case to future work.

6 Conclusion

This paper addresses the problem of detecting cite-worthiness in text, seen as the task of flagging a missing reference to a scientific article that should come to support a claim formulated in the text. While previous work has mainly taken interest in this problem from a scientific literature perspective, in our study, we extend this idea to the social media context. The paper lays ground for a discussion as of how flagging missing scientific references in claims made on social media can help improve the quality of societal debates and increase trust in social media platforms. Our preliminary results show that state of the art models applied on scientific literature corpora perform less well when let to deal with claims coming from X/Twitter. This observation opens the way for research into the development of (a) larger annotated datasets for cite-worthiness detection on claims from X and (b) the development of language models tailored for scientific discourse that could be fine-tuned for that specific task.

Acknowledgements

This work is supported by the AI4Sci grant, co-funded by MESRI (France, grant UM-211745), BMBF (Germany, grant 01IS21086), and the French National Research Agency (ANR).

References

1. Wright, D., & Augenstein, I. (2021, August). CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 1796-1807).
2. Beltagy, I., Lo, K., & Cohan, A. (2019, November). SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3615-3620).
3. Hafid, S., Schellhammer, S., Bringay, S., Todorov, K., & Dietze, S. (2022, October). SciTweets-A Dataset and Annotation Framework for Detecting Scientific Online Discourse. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (pp. 3988-3992).
4. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.
5. Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2020, July). S2ORC: The Semantic Scholar Open Research Corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4969-4983).
6. Färber, M., Thiemann, A., & Jatowt, A. (2018). To cite, or not to cite? Detecting citation contexts in text. In Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40 (pp. 598-603). Springer International Publishing.
7. Della Giusta, M., Jaworska, S., & Vukadinović Greetham, D. (2021). Expert communication on Twitter: Comparing economists' and scientists' social networks, topics and communicative styles. Public understanding of science, 30(1), 75-90.
8. Moturu, S. T., & Liu, H. (2011). Quantifying the trustworthiness of social media content. Distributed and Parallel Databases, 29, 239-260.
9. Sundriyal, M., Akhtar, M. S., & Chakraborty, T. (2023). Leveraging Social Discourse to Measure Check-worthiness of Claims for Fact-checking. arXiv preprint arXiv:2309.09274.
10. Satapara, S., Mehta, P., Ganguly, D., & Modha, S. (2024). Fighting Fire with Fire: Adversarial Prompting to Generate a Misinformation Detection Dataset. arXiv preprint arXiv:2401.04481.
11. Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. Journal of medical Internet research, 13(4), e2012.
12. Jain, N., & Singh, M. (2021, September). TweetPap: a dataset to study the social media discourse of scientific papers. In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 328-329). IEEE.
13. August, T., Card, D., Hsieh, G., Smith, N. A., & Reinecke, K. (2020, April). Explain like I am a Scientist: The Linguistic Barriers of Entry to r/science. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-12).

14. Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., ... & Gilbert, E. (2018). The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1-25.
15. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2), 211-236.
16. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. science, 359(6380), 1146-1151.
17. Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. ACM Transactions on Social Computing, 1(1), 1-27.
18. De Semir, V. (2000). Scientific journalism: Problems and perspectives. International Microbiology, 3(2), 125-128.
19. Dunwoody, S. (2021). Science journalism: Prospects in the digital age. In Routledge handbook of public communication of science and technology (pp. 14-32). Routledge.
20. Arnold, Phoebe. (2020). The challenges of online fact checking. Technical report, Full Fact
21. Didegah, F., Mejlgaard, N., & Sørensen, M.P. (2018). Investigating the quality of interactions and public engagement around scientific papers on Twitter. J. Informetrics, 12, 960-971.
22. Liu, Y., Whitfield, C., Zhang, T., Hauser, A., Reynolds, T., & Anwar, M. (2021). Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. Health Information Science and Systems, 9.
23. Raza, H., Faizan, M., Hamza, A., Mushtaq, A., & Akhtar, N. (2019). Scientific Text Sentiment Analysis using Machine Learning Techniques. International Journal of Advanced Computer Science and Applications.
24. Sugiyama, K., Kumar, T., Kan, M., & Tripathi, R.C. (2010). Identifying citing sentences in research papers using supervised learning. 2010 International Conference on Information Retrieval & Knowledge Management (CAMP), 67-72.
25. Bird, S., Dale, R., Dorr, B., Gibson, B.R., Joseph, M.T., Kan, M., Lee, D., Powley, B., Radev, D.R., & Tan, Y.F. (2008). The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. International Conference on Language Resources and Evaluation.
26. Färber, M., Thiemann, A., & Jatowt, A. (2018). A High-Quality Gold Standard for Citation-based Tasks. International Conference on Language Resources and Evaluation.
27. Alperin, J. P., Fleerackers, A., Riedlinger, M., & Haustein, S. (2024). Second-order citations in Altmetrics: A case study analyzing the audiences of COVID-19 research in the news and on social media. Quantitative Science Studies, 1-28.
28. Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Míguez, R., Caselli, T., ... & Kartal, Y. S. (2022). Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In 2022 Conference and Labs of the Evaluation Forum, CLEF 2022 (pp. 368-392). CEUR Workshop Proceedings (CEUR-WS. org).
29. Alam, F., Barrón-Cedeño, A., Cheema, G. S., Hakimov, S., Hasanain, M., Li, C., ... & Nakov, P. (2023). Overview of the CLEF-2023 CheckThat! lab task 1 on checkworthiness in multimodal and multigenre content. Working Notes of CLEF.
30. Fang, Z., Costas, R., Tian, W., Wang, X., & Wouters, P. (2020). An extensive analysis of the presence of altmetric data for Web of Science publications across subject fields and research topics. Scientometrics, 124(3), 2519-2549.