Detecting Stance in Scientific Papers: Did we get more Negative Recently?

Dominik Beese¹, Begüm Altunbaş², Görkem Güzeler², Steffen Eger¹
¹ Technische Universität Darmstadt, ² Sabancı University

Abstract

In this paper, we classify scientific articles in the domain of natural language processing (NLP) and machine learning (ML) into whether (i) they extend the current state-of-the-art by introduction of novel techniques which beat existing models or whether (ii) they mainly criticize the existing state-of-the-art, i.e., that it is deficient with respect to some property (e.g., wrong evaluation, wrong datasets, misleading task specification). We refer to contributions under (i) as having a “positive stance” and contributions under (ii) as having a “negative stance” to related work. We annotate over 2k papers from NLP and ML to train a SciBERT based model to automatically predict the stance of a paper based on its title and abstract. We then analyze large-scale trends on over 41k papers from the last ∼35 years in NLP and ML, finding that papers have gotten substantially more positive over time, but negative papers also got more negative and we observe considerably more negative papers in recent years. Negative papers are also more influential in terms of citations they receive.

1 Introduction

Deep learning has revolutionized machine learning (ML) and natural language processing (NLP) in the last decade. In particular, deep learning has led to unprecedented performance gains on a large number of NLP and ML tasks, including machine translation (Lample et al., 2018), image classification (Krizhevsky et al., 2012), natural language understanding (Devlin et al., 2018), and text generation (Radford et al., 2019).

On the other hand, there has seemingly also been a recent flood of papers highlighting limitations of (deep learning) approaches, including claims about models exploiting dataset biases (Niven and Kao, 2019), flawed evaluation (Marie et al., 2021),

and general “troubling trends” in machine learning practice (Lipton and Steinhardt, 2019).

Indeed, from a historic perspective, deep learning – formerly known under the name of ‘artificial neural networks’ – is a prime exemplar of a technology that has received very mixed assessments over time, ranging from initial hype to negative and positive evaluation in repeating cycles (Hendler, 2008; Guo et al., 2020).

Motivated especially by (recent) observations of negative assessment of individual papers regarding the existing literature and its claims, we define a new NLP task of determining the stance of a paper with respect to its related work.¹ We take a prototypical paper of positive stance to be one that generally accepts the premises of related work (even though it may identify specific – minor – issues and weaknesses), extends it and sets a new

¹This concept is related to positive/negative citations within a paper, which has been annotated in a few works, e.g., Teufel et al. (2006). Our work goes beyond individual citations and assesses the stance of the authors’ main message.
state-of-the-art. In contrast, a prototypical paper of negative stance concludes that related work is basically flawed, mistaken and potentially based on false assumptions (cf. Table 1). (Any particular paper may mix positive and negative elements, so we treat the task as a continuous regression problem.)

We hold this task important in order to be able to (i) analyze trends in science, (ii) which could potentially anticipate pessimistic future developments (“the end of the party”). (iii) By comparing trends across two disciplines – ML and NLP – we can also contrast the evolution and current state of each. The task is also timely, as interest in the analysis of scientific literature in the NLP community (cf. Section 2) has been steadily on the rise recently.

Our contributions are:
• we define the new task of stance detection for scientific literature;
• we provide a human annotated dataset of over 2k scientific papers, labeled for their stance;
• we provide a large-scale trend analysis on over 41k papers from the NLP and ML community in the past ∼35 years;
• we address various trend questions including (a) whether negativity has recently increased; (b) whether positive/negative papers are more influential and (c) more likely to be accepted.

We point out that ‘negativity’, as is a focus of our work, plays a central role in various contexts: in the social sciences, signed social networks are networks in which agents have positive and negative relations to each other, potentially explaining phenomena such as long-term disagreement (Altafini, 2012; Eger, 2016); in science, negative citations may be a form of self-correction (Bordignon, 2020) and publishing negative results may reduce waste on resources for disputed approaches (Mlinarić et al., 2017); in economics, the principle of creative destruction (Aghion and Howitt, 1990) may explain the evolution and progress of capitalism.

2 Related work

Historically, analysis of scientific literature is the scope of the fields of scientometrics and science-of-science. Classical results include the relation between title length and the number of citations a paper receives (Letchford et al., 2015) as well as quantitative laws underlying citation patterns or the number of co-authors of papers over time (Fortunato et al., 2018). In recent years, with the rise in quality of models and approaches, more and more NLP approaches are also devoted to the analysis of scientific literature.

Gao et al. (2019) ask how much the author response in the “rebuttal phase” of the peer review process influences the final scores of a reviewer, finding its impact to be marginal, especially compared to the scores of other reviewers.

Prabhakaran et al. (2016) predict whether a scientific topic will rise or fall in popularity based on how authors frame the topics in their work. They use a subset of the Web of Science Core Collection with papers from 1991 to 2010 and analyze abstracts by assigning scientific topics (e.g., stem cell research) as well as rhetorical roles (i.e., scientific background, methods used, results, etc.) to phrases. Prabhakaran et al. (2016) find that topics that are currently discussed as results and background are at their peak and tend to fall in popularity in the future, whereas topics that are mentioned as methods or conclusions tend to start to rise in popularity.

Arguably the paper most similar to ours is that of Jurgens et al. (2018). They study the entire content of a scientific publication in order to predict its future impact, based on how citations are framed. They differentiate between different functions of citations: BACKGROUND (the other work provides relevant information), USES (usage of data, methods, etc. from the other work), and COMPARISON OR CONTRAST (express similarities/differences to the other work). Using human annotation and trained classifiers on top of them, they annotate 21k papers from the ACL Anthology and analyze the evolution of the functions over (a) the course of a paper, (b) different venues, and (c) time. For (c), papers citing many other technologies with USES citations tend to have large future impact. They further show that the field of NLP has shown
a significant increase in consensus when authors started to use less COMPARISON OR CONTRAST citations and simply acknowledged previous work as BACKGROUND. The amount of USES citations also increased when researchers began to agree on the same methods and technologies. The authors argue that these trends imply that NLP has become a rapid discovery science (Collins, 1994), i.e., a particular shift a scientific field can undergo when it reaches a high level of consensus on its research topics, methods, and technologies, and then starts to continually improve on each other’s methods. A number of recent papers also leverages or studies citation context, including Cohan et al. (2019); Jebari et al. (2021); Wright and Augenstein (2021); Lauscher et al. (2021). Our approach differs from Jurgens et al. (2018) in several ways: for example, we do not analyze individual citations, but directly evaluate the stance of a complete paper (as measured by its framing in the paper’s abstract); most importantly, we are particularly interested in negative stances, which as relation is absent in the classification scheme of Jurgens et al. (2018).

Beyond classification, Yuan et al. (2021) use NLP models to automatically generate reviews for scientific papers. To do so, they compile a dataset of scientific articles and reviews from OpenReview and NeurIPS Proceedings. They conclude that their review generation model is not good enough to fully automate the reviewing process, but could still make the reviewer’s job more effective. Wang et al. (2020) create automatic reviews for papers from OpenReview, NeurIPS, and the PeerRead dataset (Kang et al., 2018a) by defining multiple knowledge graphs, one extracted from the paper, one from the papers the paper cites, and one for background knowledge. Beltagy et al. (2019) train a language model (extending BERT) on a large multi-domain corpus of scientific publications.

### 3 Data

We extract our data from two sources: the ACL Anthology which contains papers and meta-data for all major NLP events. Our second data source are machine learning conferences.

#### NLP dataset

From the ACL Anthology, we extract papers from eight different NLP venues between 1984 and 2021. For all venues, we only include papers from the main conference and exclude papers from workshops (by manually selecting the volumes) and contributions like book reviews and title indices (by filtering the titles). To extract the data, we download the provided metadata from the ACL Anthology website in the BibTeX format that contains information of authors, title, venue and year. We then use Allen AI’s Science Parse to extract abstract information and collect author and citation information from Semantic Scholar.

| Venue   | # papers |
|---------|----------|
| ACL     | 6,819    |
| EMNLP   | 5,013    |
| COLING  | 4,653    |
| NAACL   | 2,636    |
| SemEval | 2,384    |
| CoNLL   | 1,033    |
| CL      | 952      |
| TACL    | 351      |
| **Overall** | **23,841** |

Table 2: NLP dataset.

#### ML dataset

We download papers from the respective websites of NeurIPS, AAAI, ICML, and ICLR, and then use Science Parse to extract abstracts and Semantic Scholar to collect citation information as above. ML contains over 18k papers between 1989 and 2021. The distribution over the venues is given in Table 3.

| Venue   | # papers |
|---------|----------|
| NeurIPS | 7,537    |
| AAAI    | 4,384    |
| ICML    | 3,667    |
| ICLR    | 2,720    |
| **Overall** | **18,308** |

Table 3: ML dataset.

### 4 Data annotation

We annotate the data from NLP and ML for each paper’s stance towards related work (as given in the authors’ framing in a paper’s abstract). In contrast to some related work, we do not annotate stance towards individual citations, but infer the
authors’ stance from the paper abstracts and titles as we are interested in the stance of the authors’ overall core message. Indeed, in abstracts, authors typically condense the most important information they intend to convey. This agrees with the insight that abstracts and titles are typically the only piece of the paper that the majority of readers consumes (Andrade, 2011). Our focus on title and abstract is therefore a deliberate choice.

On a coarse-grained level, we consider three possible stances: (1) a paper may express a positive stance towards related work, e.g., extend the current approaches and reach a new state-of-the-art; (2) a paper may express a negative stance towards related work, e.g., state that datasets, evaluation protocols or techniques are basically flawed; (3) if a paper does not particularly fit into this categorization, e.g., because it discusses or summarizes previous work without criticizing it, then we consider a paper as expressing a neutral stance. In the annotation, we relax this coarse-grained scheme and instead give continuous numbers for annotation, ranging from -1 (severe negative stance) towards +1 (very positive stance), with 0 as neutral stance. We provide guidelines for the annotation in the Appendix.

In the following, we describe the annotation process and provide statistics.

4.1 Annotation statistics and procedure

In total, we manually annotated 2,170 papers from NLP and ML. The distribution of paper over venues is given in Table 4. In this human annotated dataset, the ACL conference, which takes place annually since 1979, has most papers (353), followed by NeurIPS (246) and AAAI (218). We also included some workshop papers (WS) in our human annotated, but remove them in the automatic analysis later on, except for SemEval.

| Venue | # papers | Venue | # papers |
|-------|----------|-------|----------|
| ACL   | 353      | EMNLP | 180      |
| NeurIPS | 246    | COLING | 151     |
| AAAI  | 218      | WS    | 87       |
| NAACL | 207      | CoNLL | 67       |
| ICLR  | 200      | SemEval | 50     |
| ICML  | 200      | Other | 21       |
| CL    | 190      |       |          |

Table 4: Human annotated dataset.

In our human annotated data, 1,338 papers (61.7%) exhibit positive stance (≥ 0.1), 389 papers (17.9%) exhibit negative stance (≤ −0.1), and 443 papers (20.4%) are neutral (> −0.1 ∧ < 0.1). We show the more detailed distribution of papers in terms of stance in Figure 1.

This statistic does not reflect the true distribution of stance in our data (which is dominated by positive papers, as we will show below), as we oversampled negative papers using heuristics (e.g., looking for particular keywords in abstracts and titles such as fail, limitation). We did this in order to ensure classifiers trained on datasets that are not too class imbalanced.

Inter-annotator agreement We had up to 5 annotators annotate abstracts for stances. The annotators were computer science undergraduate students and one computer science faculty member from NLP. 71% of the human annotated data set is annotated by one, 22% by two, 5% by three, and 2% by four annotators. We label each abstract’s stance as the average over all the annotators.

We measure agreement on stance annotation using Pearson’s Correlation Coefficient as well as Cohen’s Kappa Coefficient. The resulting correlations among all pairs of annotators (on a common set of instances) range from 0.67 to 0.95. On a coarse level with three stances, we have an agreement between 0.56 to 0.86 kappa. Overall, we thus have good agreement.

Historic vs. modern NLP data We refer to NLP papers published before the year 2000 as historic papers (Hist) and papers published since 2000 as modern papers (NLP). The historic dataset consists of 350 papers, the modern dataset of 1,020 papers.

5 Model implementation

For all our experiments, we use SciBERT (Beltagy et al., 2019). We feed each paper as concatenation of title and abstract separated by special tokens to SciBERT: [CLS] <title> [SEP] <abstract>. We set the maximum token length to 300, which is sufficient for most papers and a good compromise for efficient memory usage. We add a fully connected layer with one output neuron and linear activation on top of the pooled output to obtain a single prediction for the stance of a paper. Since we define stance as a value between -1 and +1, we clip the prediction to the desired range.

The model is trained with the following hyperparameters, which we found to work best in a prelim-
inary experiment by performing a full grid search over the hyperparameters given in parenthesis and evaluating on the dev set: a batch size of 16 (8, 16, 32), a slanted triangular learning rate (Howard and Ruder, 2018) with a maximum learning rate of 5e-5 (5e-6, 1e-5, 2e-5, 5e-5, 8e-5, 1e-4), a warmup ratio of 0.06, and linear decay. We train for 3 epochs (2, 3, 4, 5) and optimize using Adam (Kingma and Ba, 2015) with an $\epsilon$ of 1e-6, $\beta_1$ of 0.9, $\beta_2$ of 0.999, and the MSE as the loss function.

For our experiments in Section 6, we train 10 models with those hyperparameters and keep the best model based on the MSE on the dev set. We repeat this 3 times and calculate performance metrics (cf. Section 6.1) based on those models.

6 Experiments

In the following, we first verify the reliability of our stance detection model described in Section 5. To do so, we assess its cross-domain and in-domain performance and compare it to several baselines. Once the quality of the model is assured, we apply it large-scale to determine trends over time and venues in Section 7.

6.1 Experimental setup

**Metrics** We use two different metrics to evaluate our models. The Coefficient of Determination ($R^2$) is similar to the Mean Squared Error (MSE) but also takes the distribution of the data into account which makes it more informative and truthful than the MSE (Chicco et al., 2021). A model that always predicts the expected value has an $R^2$ score of 0. The range of the metric is $(-\infty, 1]$. The Macro F1 Score (M. F1) is a standard metric to assess the quality of multi-class classification which can take class-imbalance into account. We compute the macro F1 score on coarse-grained stance labels (positive, negative, neutral), see above.

**Baselines** We compare our models to simple baselines. POS: Always predict a positive stance +1; ZERO: Always predict a neutral stance 0; NEG: Always predict a negative stance -1; AVG: Always predict the average of manual annotations.

**Cross-domain experiments** Due to exponential growth of science, our model is mostly trained on more recent data. However, we also want to make sure that we obtain reliable predictions, e.g., for past papers. As a consequence, we first evaluate our model in a cross-domain setting. In this, we train our model on papers from different time stamps or domains and evaluate on a respective out-of-domain test set. For the source data, we set the train-dev-split ratio to 0.7 and 0.3, and we use the whole annotated data for the target data.

**In-domain experiments** We also perform in-domain tests as an upper bound for our cross-domain evaluation. We set the train-dev-test ratio to 0.6/0.1/0.3.

**Combined experiments** We create a combined test set which consists of NLP, ML and Hist papers. We set the train-dev-test ratio to 0.6/0.1/0.3.

**Data** As mentioned in Section 4.1, we divided our human-annotated dataset into 3 groups. The Hist dataset consists of a total of 350 papers, of which 14% negative, 44% are neutral, and 42% positive. The NLP dataset consists of 1,020 annotated abstracts and has 21% negative, 17% neutral, 62% positive papers. The ML dataset consists of 800 annotated samples and has 15% negative, 14% neutral, 71% positive papers.

6.2 Results

Results are shown in Figure 2. We observe a very clear trend: the in-domain performance is typically better than the cross-domain performance which is
better than the (best) baselines. The model trained on combined data even outperforms the in-domain results. In-domain and cross-domain performance on Hist is lowest, which is not surprising as this dataset is smallest in size and, due to temporal divergence (Lazaridou et al., 2021), assumedly has largest divergence to the modern datasets. In general, however, cross-domain performance does not lag too much behind the other experimental conditions, suggesting that our model is reliable across (our) domains. It is interesting to note that a model trained on a combination of all time periods performs best. It even leads to good results on the Hist portion of our data, which is why we use it for the analysis below.

In Table 5 (in the appendix), we illustrate sample predictions of our best performing model, randomly sampled from a predicted stance of 1.0 (very positive papers) or below -0.8 (very negative papers). We note that the predicted values look very plausible, except possibly for the very positive paper from ML, which also contains some negative elements and should probably be labeled as slightly less positive.

7 Analysis

We analyze large-scale trends from the model’s predictions and smooth graphs with Gaussian blurs.

“Are there more positive or more negative papers?” The histogram of stance values predicted by the model, aggregated over all venues and years, is visualized in Figure 3a. It shows that most papers have a positive stance and that the more negative the stance gets, the fewer papers there are. Less than 4% of all papers have a negative stance and more than 80% of all papers have a stance of ≥0.6.

“Are NLP papers more positive/negative than ML papers?” Figure 3b shows the histogram of the predicted stance values, aggregated over all years, for both datasets. The distribution is similar, but ML has more papers with stance values between 0.5 and 0.8, whereas NLP has more papers with a stance of 1.0. Overall, 4.3% of all NLP papers and 2.1% of all ML papers exhibit a negative stance, which makes NLP more negative than ML.

“Did we get more positive/negative recently?” We analyze the development of the average stance value over time in Figure 4. This shows that the average stance is always positive with a minimum value of 0.29 for NLP in the 1980s. When our ML dataset started in 1989 it was more positive than NLP. In the late 1990s, the stance of papers in NLP and ML came closer together, but around 2000 (when NAACL and SemEval were first held) NLP took over and became slightly more positive than ML. The positiveness reached its peak around 2015 with an average stance above 0.80 for NLP. It then started to get more negative for NLP which means that the field of NLP got more negative recently. The ML domain, however, got more positive with a maximum stance of 0.80 in 2021.

We further analyze whether the increase in positiveness from 1990 to 2010 and the decrease in positiveness in the most recent years for NLP comes from more positive/negative papers overall or from positive/negative papers getting less/more positive/negative. Figure 5 shows how many papers have a negative stance in each year. We observe
that negativity has peaked in the late 1980s and late 1990s for NLP and ML, respectively. There was then a continuous downward trend in negativity until the early (ML) and mid (NLP) 2010s. In the recent years, negativity has increased for both domains, but considerably more sharply so for NLP, from below 4% of all papers in 2015 to over 6%.

Figure 6 shows the average stance value of all positive papers and the average stance value of all negative papers over time. The development of the average positive stance value is very similar to the development of the average stance value overall. The average negative stance value has a decreasing trend for both datasets which means that negative papers have gotten more negative over time.

Finally, we analyze trend curves for individual venues, visualized in Figure 7. The trend towards more negative papers in the most recent years is visible for most venues, especially ACL, EMNLP, COLING, NAACL, CoNLL, ICML and ICLR. TACL has the sharpest increase. NeurIPS, AAAI, CL, and SemEval do not follow this trend, however. Many venues were more negative before the 2000s and least negative in the 2000s. The CL journal is a noticeable outlier: it is the most negative venue with up to 40% negative papers (we note that it is also the only journal in our dataset).

“Do positive/negative papers receive more/fewer citations?” Figure 8 shows how many citations a paper with a certain stance value has received in comparison to papers published in the same year. Positive values
Figure 6: Distribution of the average stance value of papers with a positive (left) and a negative (right) stance over time for both domains, average stance and 95% confidence interval. The values indicate the average positiveness/negativeness of positive/negative papers.

Figure 7: Percentage of negative papers for each venue over the years on a logarithmic scale.

Figure 8: Normalized number of citations a paper with a certain stance value has received, average number of normalized citations and 95% confidence interval. Normalized values indicate how many citations more or less a paper has received in comparison to the average number of citations a paper published in the same year has received, measured in multiples of the standard deviation of citation counts in that year.

“Do positive/negative papers have lower/higher acceptance chances?” We use 2,606 accepted and 6,100 rejected papers from the ICLR conferences in 2013 and 2017–2021, collected from OpenReview, and 3,063 accepted and 9,142 rejected papers from ACL, EMNLP, NAACL, NeurIPS, AAAI, ICML, and ICLR over the years 2007–2017 as collected by Kang et al. (2018b). Figure 10 shows how many papers with a certain stance value were accepted. Although the variance indicate more citations than the average, negative values indicate fewer citations. The graph shows that papers with a negative stance receive more citations than the average paper in the same year; very negative papers receive even more citations; except for the most negative papers with stance values of -1.0. In contrast, a paper with a positive stance receives less citations on average, but very positive papers with a stance value of 1.0 receive slightly more citations.

Similar results can be found for NLP by analyzing the individual domains separately, which is shown in Figure 9. The domain of ML is more extreme in that negative papers receive even more citations and positive papers with stance values from 0.1 to 0.8 even less.
consistent with the trend over time, which shows that fewer papers were negative in the years 2007–2014 than in 2015–2021, implying a bias: positive papers were more popular back then and therefore more positive papers got accepted.

We also calculate the acceptance rates for two separate time spans, 2007–2014 and 2015–2021, as shown in Figure 11. For the most recent years 2015–2021, the trend is similar to the overall trend with the exception that very positive papers have higher acceptance rates. The acceptance rate for papers with stance values between -0.6 and 0.8 is lower than the overall acceptance rate, but for very positive papers the acceptance rate is slightly higher again.

Figure 11: Normalized acceptance rate of a submitted paper with a certain stance value for two time spans, average acceptance rate and 95% confidence interval. Normalized values indicate how many percentage points more or less a paper with a certain stance value is likely to be accepted in comparison to the average acceptance rate in each time span. The dotted line indicates the average acceptance rate.

8 Concluding remarks

We analyzed stance in abstracts of scientific publications, where authors position themselves positively or negatively with respect to related work. We annotated over 2k abstracts from ML and NLP venues and trained a SciBERT model on a subset of the annotated abstracts, verifying that the model is of sufficiently high quality for the task. We then used this model to automatically predict the stance of a paper based on its title and abstract. We applied the model large-scale to a collection of 41k scientific publications in the domain of NLP and ML from the years 1984 to 2021 to enable large-scale analysis.

The analysis revealed that the majority of papers in the past and today have a positive stance, that the average stance has substantially increased over time, yielding support for the hypothesis that ML and NLP have become “rapid discovery sciences”, and that the ML domain is more positive than the NLP domain. Scientific publications used to be more negative in the early days, then became
very positive until they started to get more negative again recently. Overall, publications got also more extreme over time, which means that papers with a positive stance became more positive and papers with a negative stance more negative. We found negative papers to be more influential than positive ones in terms of citations they receive and, recently, more likely to be accepted to NLP/ML venues.

Future work should address other scientific disciplines beyond NLP and ML for a broader scientific trend analysis, examine the correlation between overall stance of a paper and individual (negative) citations in its related work sections, annotate word-level rationales for our sentence-level scores, assess the correlation between stance and socio-demographic factors (gender, nationality, affiliation, h-index, etc.) and analyze how negative papers may potentially transform a field.

References
Philippe Aghion and Peter Howitt. 1990. A model of growth through creative destruction.
Claudio Altafini. 2012. Consensus problems on networks with antagonistic interactions. *IEEE transactions on automatic control*, 58(4):935–946.
Chittaranjan Andrade. 2011. How to write a good abstract for a scientific paper or conference presentation. *Indian Journal of Psychiatry*, 53:172 – 175.
Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. *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), pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
Frederique Bordignon. 2020. Self-correction of science: a comparative study of negative citations and post-publication peer review. *Scientometrics*, 124:1225–1239.
Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. *Bidirectional recurrent convolutional neural network for relation classification*. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 756–765, Berlin, Germany. Association for Computational Linguistics.
Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. 2021. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623.
Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. *Structural scaffolds for citation intent classification in scientific publications*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.
Randall Collins. 1994. *Why the social sciences won’t become high-consensus, rapid-discovery science*. *Sociological Forum*, 9(2):155–177.
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Steffen Eger. 2016. Opinion dynamics and wisdom under out-group discrimination. *Mathematical Social Sciences*, 80:97–107.
Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science*, 359(6379).
Yang Gao, Steffen Eger, Illia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1274–1290, Minneapolis, Minnesota. Association for Computational Linguistics.
Qinghua Guo, Shichao Jin, Min Li, Qiuli Yang, Kexin Xu, Yuanzhen Ju, Jing Zhang, Jing Xuan, Jin Liu, Yanjun Su, et al. 2020. Application of deep learning in ecological resource research:
Theories, methods, and challenges. *Science China Earth Sciences*, pages 1–18.

James Hendler. 2008. Avoiding another ai winter. *IEEE Intelligent Systems*, 23(2):2–4.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Chaker Jebari, Enrique Herrera-Viedma, and Manuel Jesus Cobo. 2021. The use of citation context to detect the evolution of research topics: a large-scale analysis. *Scientometrics*, 126(4):2971–2989.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018a. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018b. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, David Jurgens, Arman Cohan, and Kyle Lo. 2021. Multicite: Modeling realistic citations requires moving beyond the single-sentence single-label setting.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d’Autume, Tomáš Kočíský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Adrian Letchford, Helen Susannah Moat, and Tobias Preis. 2015. The advantage of short paper titles. *Royal Society open science*, 2(8):150266.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77.

Q. Liu, Shu Wu, and Liang Wang. 2015. Cot: Contextual operating tensor for context-aware recommender systems. In *AAAI*.

Benjamin Marie, Atsushi Fujita, and Raphael Rubinio. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769
Appendix

Annotation guidelines

We issued the following guidelines to annotators: (1) The stance is a value in the range from -1 (very negative) to +1 (very positive). (2) Only the title and abstract of a paper is taken into account. A contribution has a positive stance and is annotated with a positive number up to +1 when: (a) it clearly indicates to improve the state-of-the-art by beating existing standards; (b) it presents novel techniques; (c) it proposes solutions to problems of previous work; (d) it gives insights to existing models or methods and explains why they work. A contribution has a negative stance and is annotated with a negative number up to -1 when: (a) it clearly criticizes previous work of being wrong; (b) it presents flaws of existing work, i.e. that an approach is deficient with respect to some property; (c) it analyzes errors of other methods and explains why they do not work as expected. Contributions that have positive and negative parts are annotated with a value between -1 and +1, taking into account the following: (a) The importance of individual parts matters, i.e., ‘some problems’ is less negative than ‘fails to work’. (b) The amount of positive and negative parts matters. (c) The last sentence of an abstract is usually the most important sentence. If the last sentence of a contribution is positive, it is more positive than a contribution with a negative last sentence. Contributions that fall outside this labeling scheme are neutral and annotated with a 0. Those include: (a) contributions that explore existing work without beating other systems or explaining why it works or doesn’t work; (b) contributions that compare, discuss, study, or summarize existing work without criticizing it.
In this paper, we present a novel model BRCNN to classify the relation of two entities in a sentence. [...] We propose a bidirectional architecture to learn relation representations with directional information along the SDP forwards and backwards at the same time [...] Experimental results show that our method outperforms the state-of-the-art approaches [...]” (Cai et al., 2016)

However, the state-of-the-art context modeling methods treat contexts as other dimensions similar to the dimensions of users and items, and cannot capture the special semantic operation of contexts. [...] In this work, we propose Contextual Operating Tensor (COT) model, which represents the common semantic effects of contexts as a contextual operating tensor and represents a context as a latent vector. [...] Experimental results show that the proposed COT model yields significant improvements over the competitive compared methods on three typical datasets [...]” (Liu et al., 2015)

Distributional representations of words have been recently used in supervised settings for recognizing lexical inference relations between word pairs, such as hypernymy and entailment. We investigate a collection of these state-of-the-art methods, and show that they do not actually learn a relation between two words. Instead, they learn an independent property of a single word in the pair: whether that word is a ‘prototypical hypernym’.” (Levy et al., 2015)

In this paper, we investigate whether state-of-the-art MRC models are able to correctly process Semantics Altering Modifications (SAM): [...] We present a method to automatically generate and align challenge sets featuring original and altered examples. We further propose a novel evaluation methodology to correctly assess the capability of MRC systems [...] We [...] find that [...] optimised models consistently struggle to correctly process semantically altered data.” (Schlegel et al., 2021)

Table 5: Predicted highly positive and highly negative papers from our datasets. Blue/red text represents positive/negative rationales; these are added by us.