Trust me, I am an Expert: Predicting the Credibility of Experts for Statements

Markus Nilles[1,*], Lorik Dumani[1], Björn Metzler[1] and Ralf Schenkel[1]

[1]Trier University, Behringstraße 21, D-54286 Trier, Germany

Abstract

Nowadays, information on any topic can be researched on the Internet. However, in addition to reputable news sources, there is also a great deal of fake news that is disseminated, e.g., via social media or in established newspapers. Thus, the veracity must be assessed for each piece of information. People, parties, and organizations want to push through their interests and sometimes do not hesitate to spread fake news. For some time now, one popular means has been to quote (supposed) experts in a field. For example, —due to his authority— Albert Einstein is often quoted by believers in God although he was primarily concerned with physics while his quotes in God are taken out of context.

In this paper, we define a new task of expert suitability prediction and evaluate methods to assess the credibility of a person with reference to a statement and its context and compare it to state-of-the-art approaches applying transformer-based embeddings. In an R4 cycle in CBR this approach could be used for the ranking. In this pilot study, we restrict our experiments to researchers, which allows us to derive their expertises from their publications. Furthermore, we make a manually labeled dataset consisting of 1,700 (statement,expert) pairs where suitable experts were tediously searched out together with valuable context information (such as convincing text parts of the experts' contexts towards a statement) publicly available to stimulate further research in this very important, but up to now underrepresented area of fake news detection.

Keywords

Fake News, Expert Validation, Claim Validation, Argumentation

1. Introduction

The Internet offers countless opportunities to consume information, and social media such as Facebook and Twitter increase the likelihood of contact with news [1]. However, the consumption of news is not always consciously selected, but users come across suggested articles because they have been shared by their contacts, for example. Random access to news causes the number of report sources to increase. At the same time, the potential for encountering misinformation or disinformation increases as well. Misinformation describes unintentionally misinterpreted information which is thus spread due to a lack of knowledge. In contrast, disinformation consists of news reports created with the intention of spreading false information. This makes it particu-

ICCBR TMG'23: Workshop on Text Mining and Generation at ICCBR2023, July 17 – 20, 2023, Aberdeen, Scotland

*Corresponding author.

nillesm@uni-trier.de (M. Nilles); dumani@uni-trier.de (L. Dumani); s4bjmetz@uni-trier.de (B. Metzler); schenkel@uni-trier.de (R. Schenkel)

0000-0002-3449-9319 (M. Nilles); 0000-0001-9567-1699 (L. Dumani); 0000-0003-3346-9844 (B. Metzler); 0000-0001-5379-5191 (R. Schenkel)

CEUR Workshop Proceedings (CEUR-WS.org)

larly important to recognize fake news, which includes both misinformation and disinformation, and thus distinguish it from true facts. Misleading information has far-reaching consequences and the global economic damage is estimated at 78 billion US dollars annually [2]. This affects such industries as politics, finance and the advertising sector. Especially with complex topics such as corona viruses, it is difficult to distinguish a serious report from false news not only for laymen.

Due to its effectiveness, a frequently used tactic to strengthen the message of own views is to quote scientists on topics that are not the core area of their research. An infamous example is Albert Einstein, who is often quoted by believers in God, although Einstein was primarily concerned with physics and his quotes in God are taken out of context [3]. In other words, people misuse the authority of an expert in order to increase the trustworthiness of another matter, while the expert's expertise is rather to be found somewhere else. Regarding this problem, it is not even important whether the expert is aware that they has been used for a statement of another field, or whether they do this themselves. Accordingly, due to the ever increasing and faster dissemination of information it is important to have a system that checks whether the person making a statement or being cited for a statement also has the expertise with regard to the topic of the statement to believe it and not only a lot of authority.

To tackle this issue, this paper presents a pilot study for validating the suitability of experts towards statements. To the best of our knowledge, there is no prior work addressing this. Since it is harsh and controversial in practice to generalize when a person is an expert in a field, we restrict our experiments to researchers only, since (1) we can derive their expertises from their publications and (2) the expertise of scientists with experience in a field is usually not disputed; the extension of the application to non-scientists is then part of the future work. Certainly, people who neither conduct research on a topic nor have published on it can also be experts on a number of subjects but they are usually not consulted as experts on controversial matters to increase trust such as energy supply or behavior in the event of military action. Thus, in this paper we pursue to answer the question whether somebody would believe a supposed expert if they made a certain statement. However, to find out whether the expert in question actually made the statement is not part of this work. In CBR, this approach could be used as ranking component in the R4 cycle.

We make the following contributions:

- **(i)** We define the new task of assessing researchers' expertises regarding controversial statements.
- **(ii)** We provide a dataset containing 1,700 manually labeled (statement,expert) pairs along with important contextual information to push this research forward.[1]
- **(iii)** We investigate the performance among others with state-of-the-art machine-learning approaches to make predictions about researchers' expertises towards statements.

Next, we address related work in Section 2. In Section 3 we define the task of assessing researchers' expertises regarding controversial statements and introduce the dataset. Then, we present our methods and report the results of our evaluation in Section 4. We conclude the paper and give an outlook to future work in Section 6.

[1]The dataset is available at the following link: https://doi.org/10.5281/zenodo.6586678

2. Related Work

Due to the increasing prevalence of fake news, the research area of fact checking is becoming more and more prominent in NLP. However, to the best of our knowledge, no one has validated expert statements and particularly not by examining scientific publications of researchers. Hence, we briefly survey the state of research (1) in general, (2) particularly on evidence-based research studies as our approach might be used to identify resources to fact check claims, and (3) credibility prediction studies as well as (4) data fusion and worker expertise in crowdsourcing as our method has some similarities to them.

Fake News detection in general In their meta-analysis, Thorne et al. [4] summarize the current state of research of automatic fact-checking and divide claim validation into verification and fact-checking. Basically, there are two concepts of how the facts of claims are verified. While some approaches perform verification using knowledge bases [5], e.g. by using WordNet [6], others use Natural Language Inference [7] to check short sentences. The tool ClaimBuster [8] determines how check worthy each sentence of a given input is and finds similar statements in a database to assess its truthfulness. While some works rely solely on the sentence [9], others consider additional metadata such as speaker profiles [10, 11] or linguistic features [12, 13, 14, 15, 16]. However, other works address the trustworthiness of websites by evaluating the graph connections and ignore the content [17].

Evidence-Based Research Studies and Credibility Prediction Studies Fields that are related to our paper to a certain degree are, to the best of our knowledge, evidence-based research studies and credibility prediction studies.

Those works consider social networks such as Twitter and measure, e.g., user influence [18]. The work by Canini et al. [19] is closest to ours, as it ranks social network users according to their credibility to a topic by combining the analysis of the link structure of social networks with topic models of the content of messages to identify and evaluate topically relevant and credible sources of information in social networks. They define credibility as the combination of expertise and trust, and expertise as the support of other professionals [20]. Note that in our paper, this support of researchers is achieved through the acceptance of papers in a peer-reviewed process at a conference.

Source Credibility in Data Fusion and Worker Expertise in Crowdsourcing Other related fields are source credibility in data fusion and worker expertise in crowdsourcing. In the field of data fusion, the goal is to combine data from multiple sources to achieve a more accurate overall picture than if only one data source is considered. For the calculation of the overall picture, it is important to know the credibility of the individual sources and to include them in the calculation [21]. MacDonald et al. [22] modeled expert ranking as a voting problem. In the ranking, the documents voted for the expert candidates and the score of each candidate was calculated using various data fusion techniques, e.g. Reciprocal Rank. Also in crowdsourcing systems such as Amazon Mechanical Turk, in which people from different fields collaborate with each other, it is important to determine the expertise of the people in advance in order to determine the best qualified person for the task [23].

Table 1
Example of a statement with context information.

component	example from dataset
headline	*Infection Prevention*
background	*About this card: This Argumentation Card provides an overview of arguments for or against the implementation of infection prevention measures in all forms of residential care in the elderly. Healthcare professionals have laid down …*
query	*What arguments do nursing homes use for and against implementing infection prevention measures?*
stance	*In front of*
frame	*Feasibility*
cluster	*Infection prevention prevents unnecessary work*
statement	*Caring for sick people takes extra time and is a drain on the daily routine.*

3. Task Definition and Dataset

As indicated in Section 1, we aim to estimate a potential expert's domain knowledge with respect to a statement using the context of both the statement and the potential expert. In this section, we first define the new task, then we discuss the dataset we created for this purpose. Note that this construction requires two components: the first consists of statements with their context. The second consists of potential experts and their expertises on these statements.

Task Definition Given a statement s with context $c(s)$, as well as an expert e with context $c(e)$. The task is to determine whether e is a credible expert given their context $c(e)$, if making or being cited for the statement s in context $c(s)$. As we restrict our work to researchers, the context $c(e)$ will be represented by e's publications as well as the research interests of e. As in this study we only target to show that our approach is feasible and also to reduce complexity and to anticipate performance reasons, in this paper the expertise of each expert is represented by exactly one of their publications; an extension to all publications will be left for future work.[2] Figure 1 depicts the context information we will use for the remainder of the paper. More precisely, we will investigate in all combinations of contexts for both experts and statements. The contexts are split into *min*, *focused*, and *all*. Here, *min* represents the minimum that could be required as context. The set *focused* expands the set *min* with further context that is available, at least for a longer period immutable, and realistic to be used in an application. The last set *all* contains even more context with the goal to examine its impact (1) either with information where the methods to determine are still being researched as in the case with the context of the statement and (2) with data that is rather related to the authority of a researcher as in the case of an expert's publication.

[2]We argue that this is not a drawback since in this case an expert would be classified as suitable as long as at least one of their papers fits.

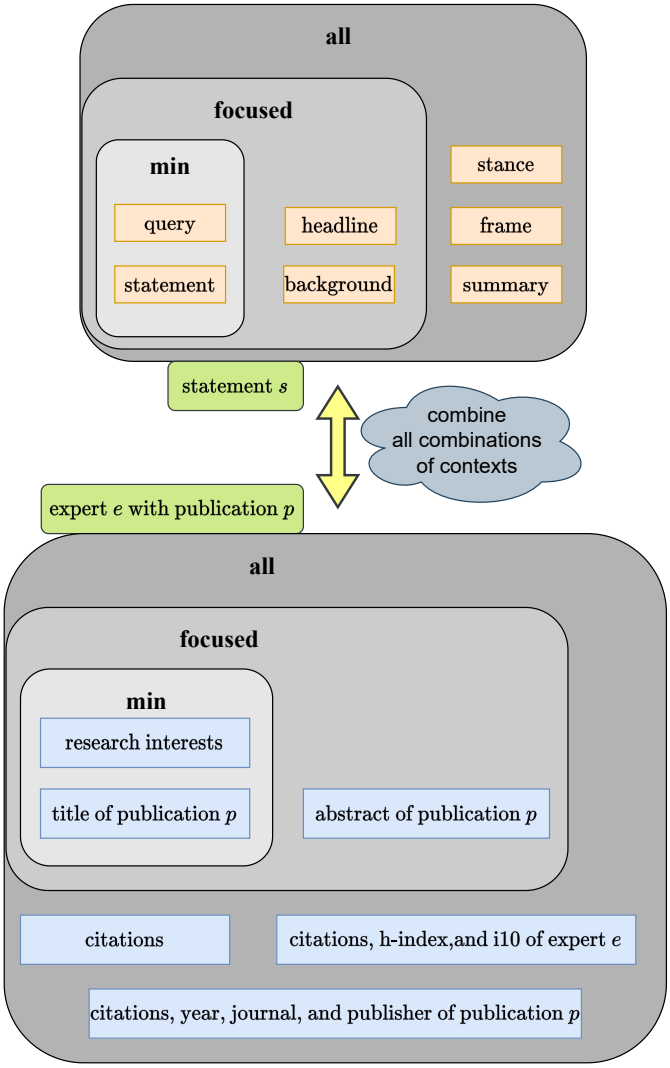

Figure 1: Visualization of the different contexts of the statement as well as the researcher being used to predict the researcher's expertise towards the statement.

Finding Suitable Statements for our Dataset We built upon the dataset from Dumani et al. [24] as a source of statements. It consists of 49 argument graphs having 2,186 arguments with binary stances as well as 133 different frames associated with them. This dataset was originally introduced to cluster arguments to different granularities such as their stances, their frames, or their quintessence. Figure 2 show an example of such a graph that is taken from that paper.

We decided to use this dataset for several reasons: first, it addresses various (Dutch) national as well as international political issues such as the environment or pensions, on which people might want to listen to expert opinions. Second, since the dataset consists of argument graphs, the root of each graph describes the core question, while each of the other nodes contain further partial information such as the stance, the frame, or a cluster (with a short summary with the

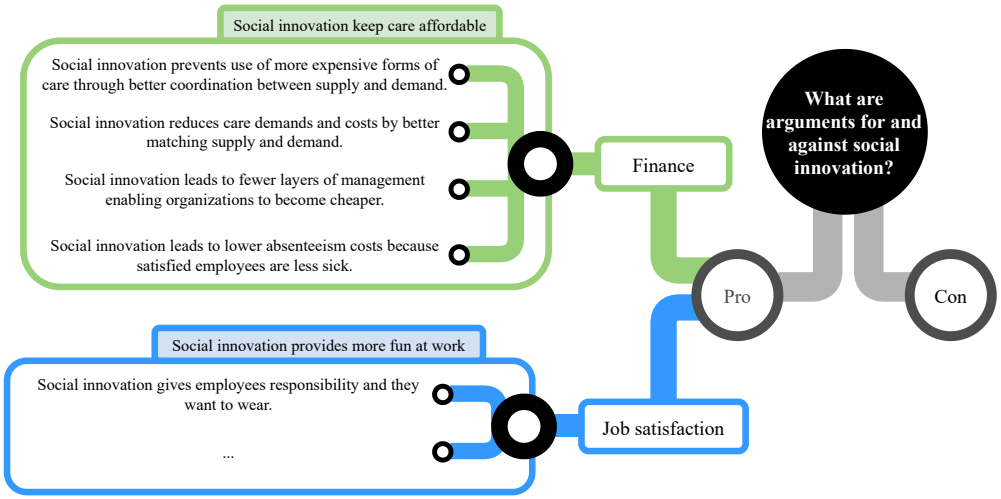

Figure 2: Example of an excerpt of an argument map (from Dumani et al. [24]).

quintessence) of the statement. As we aim to incorporate different contexts to each statement in order to evaluate the estimation of expertise, this dataset allows us to do that conveniently. Starting with this dataset, we randomly picked 200 statements with their context information to proceed from there. To avoid biases, we made the constraint that no statement would be included in this sample if there was already another one with the same stance, the same frame, and consequently the same quintessence. An example of a statement with an excerpt of its context can be seen in Table 1.

Enriching the Dataset with Suitable Experts W.r.t. these 200 statements, three members of our team searched on Google Scholar for two ideal experts for each statement, i.e. researchers who had published papers that are very related to the content of the statements. We decided to use Google Scholar as our expert source as it is the largest scientific literature database of authors and their publications including different disciplines that we are aware of. It includes not only different disciplines, but also different source documents such as professional articles, theses, dissertations, books, abstracts, or court opinions.

The exact task was to find experts for these statements who would be believed if they were cited to substantiate them. In this context, we asked the annotators to find two experts from different papers for each statement, i.e., the expert of paper 1 was not allowed to be a co-author of paper 2 and vice versa.

In a tediously task, one annotator proposed experts for statements, and the other two either approved them or replaced them with others that were in turn reviewed by the others. The annotators developed text snippets based on the statements (partially also with phrases) to retrieve suitable experts when typing them into Google-Scholar.

In finding the best variation of search snippets, they were allowed to be creative and enter words that do not appear in the sentence. The decision whether an expert was a good match or not depended on several characteristics, such as the *title* of their publications, their *abstracts*,

and their *research interests* which are indicated by the scientist .[3]

From the 200 statements, a total of 340 experts could be determined for 170 statements. The remaining could not be associated with any experts at all, as they are not suitable for verification, e.g. because the statements are far too general, or even far too specific about regional problems.

W.r.t. the example in Table 1, the statement "*Caring for sick people takes extra time and is a drain on the daily routine.*" was changed to the search snippet "*daily effects of caring for sick people*" for which Google Scholar, finds, i.a., the paper titled "*Differences in impact of long term caregiving for mentally ill older adults on the daily life of informal caregivers: a qualitative study*" [25]. In this case, the annotators listed the part "*Caregivers themselves are often aged, and although caregiving implies an impact on daily life that exceeds the boundaries of usual informal care*" of the abstract to be the persuasive part to believe in the expert's knowledge towards that statement.

Rating of Random Scientists to the Statements Based on these 170 statements and the 340 experts, we randomly picked eight additional experts from this expert pool for each statement, yielding a total of 1,360 additional (statement,expert) pairs. Two annotators independently inspected the researchers' Google Scholar profiles for these pairs and, based on the research interests, as well as their paper titles and abstracts, assigned the labels 0, 1, and 2, indicating whether the researcher is a suitable expert to make the statement. A third annotator also assigned a label if the ratings of the first two differed. This happened 532 times which shows the difficulty of this task caused by the subjectivity of the annotators.

The task was to assume that someone refers for statement s to expert e. Then, the annotators should indicate on the basis of the e's research whether they would be convinced while neglecting the truth content of s, i.e., they just concentrated on the topics of s and the expertise of e.

The label 0 was assigned if the expert did not match the statement at all. This could be, e.g., if the expert is a psychologist, but the statement deals with processors. In contrast to that, the label 2 was assigned if the expert is a perfect fit for the statement. The label 1 was assigned to experts who fit partially. This decision is slightly more difficult to assign because this is more subjective. The annotators were instructed to assign e.g. an economist the score 1, if the statement has something to do with economics. Here we made the assumption that they must have learned the same in their basic studies and are therefore reasonably familiar with each other's subjects. However, obviously, this assumption may not always be correct.

Altogether, we received 3,252 (=2 · 1,360 + 532) assessments for the 1,360 statements. Label 0 was assigned a total of 2,192 times, label 1 a total of 909 times, and label 2 another 151 times.

We measured the robustness of the annotations using Krippendorff's α. Measuring the assessments of the first two annotators resulted in an inter-annotator agreement (IAA) of 0.262 (interval metric). This IAA improved to 0.411 by adding the third annotator. According to Landies and Koch [26], values between 0.4 and 0.6 represent a moderate agreement.

This value is quite low and indicates the difficulty for this task from a completely different point of view, namely the subjectivity in evaluating expertise. We argue that the annotations are nevertheless better than a first glance would suggest as the third annotator knew that the

[3]Note that in addition to the expert itself, we include both the search snippets and the text snippets that convinced the annotator in our dataset.

two annotators had previously given unequal annotations, and the latter therefore weighed particularly carefully what to assign.

Final dataset Finally, our dataset consists of a total of 1,700 labeled (statement,expert) pairs, where 340 were labeled with a score of 2 because the associated experts were manually picked for these statements. For the remaining 1,360 pairs, we used majority voting of the three annotators for the final label. Especially in light of the improved IAA, this seemed reasonable. W.r.t. the 1,360 pairs, 997 were assigned with label 0, another 309 were assigned with label 1, and 54 were assigned with label 2.

4. Methods

With the last section introducing a dataset suitable for learning and evaluating the prediction of researcher's expertise, in this section we now present and evaluate methods for accomplishing this.

We examine three approaches: Our main approach bases on a transformer-based and fine-tuned cross-encoder model. In particular, as we want to measure the performance of different contexts (see Section 3), we consider all 9 (=3 · 3) combinations of the subsets $\{min, focused, all\} \times \{min, focused, all\}$ of contextual information as shown in Figure 1 and explained in Section 3.

The next approach serves as baseline and makes use of the state-of-the-art BERT where the resulting embeddings are classified by several, i.e., seven standard classifiers such as a support vector machine or gradient boosting.

The last approach serves as a comparison and makes use of a classical and very successful IR approach, namely BM25F [27], which is an extension of the famous BM25 method with document structure and anchor text.

Classifiers based on Cross-Encoders Due to the great success of transformer-based embedding methods, which have brought about great positive impact in the field of NLP, it is appropriate to use a state-of-the-art model for predicting the expertises of researchers with respect to statements. After weighing the advantages and disadvantages of the various frameworks and approaches, Our main approach consists of a cross-encoder [28]. Unlike bi-encoders, which produce a separate embedding for each text input and then use, for example, the cosine similarity of vectors to measure their similarity, Note, that cross-encoders generate an embedding for two simultaneously introduced texts. In our case, one text input consists of the expert's information and its context, and the other text content consists of the statement and its context. Cross-encoders usually perform better than bi-encoders, but have the disadvantage that they can only be used on predefined sets. However, since the set of potential experts is usually manageable and only needs to be updated at longer intervals, we weighed that it is reasonable to apply this approach.

In our experiments, we created 10 folds for this purpose and evaluated them via cross-validation. We employed the python framework SENTENCE-TRANSFORMERS and utilized the model "`roberta-large`". More precisely, we fine-tuned this model for each combination and

each (train,test) fold for 3 epochs always with a batch size of 16. We refer to this method using the cross-encoder with CE_{RoBERTa}.

Baseline utilizing BERT and Standard Classifiers As a baseline, we trained several standard classifiers. As input, the classifiers received embeddings created with the pre-trained Sentence-BERT model [29] "`all-roberta-large-v1`". To create the input for the classifiers, we computed an embedding vector for the expert's context and an embedding vector for a statement and its context and concatenated them. We applied the Python library SCIKIT-LEARN [30] for initializing and training the classifiers. We utilized the following algorithms in their default configuration for classification: Multi-layer Perceptron (MLP_{RoBERTa}), Nearest Neighbor (KNN_{RoBERTa}), Gaussian Naive Bayes (GNB_{RoBERTa}), Gradient Boosting (GB_{RoBERTa}), Random Forest (RF_{RoBERTa}), Support Vector Machine (SVM_{RoBERTa}) and Logistic Regression (LR_{RoBERTa}).

BM25F This approach serves as comparison. The intuition behind this is the assumption that the experts' knowledge represented by their textual publications could have textual overlap to a statement or its context. In order to implement the approach with BM25F, we first indexed all experts together with their context information (as shown in Figure 1) using the Java framework APACHE LUCENE. The goal here is to enter a statement and get a list of potentially matching experts. The main difference between BM25 and BM25F is that the latter allows us to add more fields than just one to the query.[4]

We therefore added the fields from the statement's context, which are shown in Figure 1, to the queries. For example, with the combination $min \times min$, we have a total of four search fields, since we look for both the *statement* and the *query* in the *research interests* and the *titles*. When querying the statement, the result is a list of ten experts with scores that are above APACHE LUCENE's internal and default threshold, but do not yet correspond to our labels 1 (partial expert) and 2 (full expert) and thus have to be converted. [5] In order to get the maximum out, we use wildcards for this prediction, i.e., the correct prediction 1 or 2 is automatically assigned to the expert if BM25F lists the researcher in its list. Using this oracle, we are able to detect the upper bound of this approach with more clarity. We coin this method $\text{ORACLE}_{\text{BM25F}}$.

5. Evaluation

We measured the performance with precision, recall, F_1, and accuracy. Therefore, we computed these mean average values in a 10 fold cross validation for each context combination. More precisely, the precision is calculated by the fraction of the experts that were predicted as 1 or 2 for which the prediction was correct. Further, the recall is computed by the fraction of the experts labeled as 1 or 2 for which the prediction was correct.[6] To better interpret these values, we also included a method called ZERO in the evaluation that always predicts 0 because

[4] In the Java framework APACHE LUCENE this is done by using the class `BlendedQuery` [31].

[5] Note that APACHE LUCENE allows to vary the number of maximum results but our experiments with the values 5, 10, 50, and unlimited had no impact at the final results.

[6] Since BM25F can be seen as an oracle, we mapped the values 1 and 2 to a single value to boost its performance.

Table 2
Evaluation showing the mean average precision (prec), recall (rec), F_1, and accuracy (acc) values of the 10 fold cross validation for each statement (s) and expert (e) combination. Values are sorted in descending order by their F_1 values and rounded to three digits after the decimal point.

method	context(s)	context(e)	prec	rec	F_1	acc
$CE_{RoBERTa}$	*focused*	*all*	**0.783**	0.526	**0.627**	0.760
$MLP_{RoBERTa}$	min	min	0.625	0.551	0.584	0.725
$CE_{RoBERTa}$	*all*	*focused*	0.778	0.486	0.576	0.738
$CE_{RoBERTa}$	*min*	*min*	0.732	0.513	0.573	0.735
$KNN_{RoBERTa}$	*focused*	*all*	0.693	0.494	0.572	0.725
$CE_{RoBERTa}$	*focused*	*focused*	0.687	0.447	0.536	0.721
$Oracle_{BM25F}$	*all*	*focused*	0.401	**0.721**	0.516	0.439
$CE_{RoBERTa}$	*all*	*min*	0.599	0.383	0.465	0.705
$CE_{RoBERTa}$	*all*	*all*	0.551	0.314	0.395	0.691
$CE_{RoBERTa}$	*min*	*focused*	0.532	0.303	0.371	0.687
$CE_{RoBERTa}$	*focused*	*min*	0.439	0.294	0.349	0.682
$GNB_{RoBERTa}$	*min*	*all*	0.286	0.428	0.341	0.475
$CE_{RoBERTa}$	*min*	*all*	0.483	0.280	0.335	0.674
$GB_{RoBERTa}$	*focused*	*min*	0.518	0.167	0.249	0.624
$Oracle_{BM25F}$	*all*	*min*	0.402	0.169	0.238	0.552
$RF_{RoBERTa}$	*focused*	*focused*	0.761	0.137	0.229	0.64
$SVM_{RoBERTa}$	*focused*	*min*	0.81	0.117	0.2	0.634
$LR_{RoBERTa}$	*focused*	*min*	0.527	0.084	0.144	0.602
Zero	-	-	0	0	0	**0.764**
$Oracle_{BM25F}$	{ *min*, *focused* }	{ *min*, *focused* }	0	0	0	0.586

this makes up the largest class in our dataset, allowing us to better interpret the results of the evaluation.

Table 2 shows the performance of the examined methods. Note, that we excluded *all* from the expert combinations in the method $Oracle_{BM25F}$ as it does not make any sense for textual string matching, e.g., to search for the number of citations of a researcher. The values are sorted in descending order according to the F_1 score, since the accuracy alone can be misleading. Going by the accuracy, we get the best performance by always predicting 0 (see method Zero with accuracy 0.764). However, in this case we also obtain 0 for the precision, the recall, and the F_1-score.

The second best accuracy is 0.76 and is obtained with our classifier $CE_{RoBERTa}$ with the contexts *focused* for statement and *all* for the expert. The accuracy value is almost equal to that of Zero but the other scores in the table show that $CE_{RoBERTa}$ is much more useful as its F_1 score (0.627) is also the highest, revealing that this approach also provides reasonable results for partially and full experts and not only for non-experts. We also see that while $CE_{RoBERTa}$ seems to be a precision-oriented approach (0.783) $Oracle_{BM25F}$ tends to be a recall-oriented method (0.721). W.r.t. the baseline we only show the classifiers with their best performing context combinations. Following the table, the method $MLP_{RoBERTa}$ produces the second best performance with the contexts *min*, respectively, when using the F_1 score. While the recall of $MLP_{RoBERTa}$ (0.551)

performs comparably well to the best method $\text{CE}_{\text{RoBERTa}}$(0.526), $\text{CE}_{\text{RoBERTa}}$achieves the better precision (0.783 instead of 0.625) and thus the better F_1-score (0.627 instead of 0.584).

We can infer from the table that $\text{Oracle}_{\text{BM25F}}$ performs better the more context we feed into it. Note, that the good performance of $\text{Oracle}_{\text{BM25F}}$ probably results from the fact that it works partly as an oracle. Actually, it performs very poorly when it is not fed with *all* statement's context information. The rankings also suggest that the baseline is more sensitive to the context of the statement. $\text{CE}_{\text{RoBERTa}}$ performs best when providing *focused* context information of the statement and *all* context information of the expert. However, these numbers should be treated with caution, because the performance was boosted by adding publication-independent features such as the h-index. Considering only the textual content of the publication, *all* $\times$ *focused* produces the best and most realistic result when applying cross-encoders. Almost all standard classifiers perform best when the context *focused* is chosen for the statement. Too much information probably has a negative effect here. Regarding the context of the expert, it is a bit more mixed. In particular, it is noticeable that the second best method in the table, MLP_{RoBERTa}, performs best only with minimal input in each case. We suspect that the context in the statement loses its impact due to long background information in *focused*. In the future, we consider using automatic summarization methods like T5 [32] or keyword extractors like Yake [33, 34, 35] to avoid this issue. Also, we would like to investigate how the performance behaves when trying new combinations, such as $all \setminus focused \cup min$.

6. Conclusion and Future Work

Citing putative researchers to strengthen own viewpoints is a widely used means of Fake News. In this paper, we defined the task of assessing a person's expertise towards a statement and showed successfully that we can predict whether a researcher is a partially or fully suitable expert to be cited to believe a statement by fine-tuning a state-of-the-art transformer model. We make the dataset consisting of 1,700 labeled (statement,expert) pairs together with valuable information to train search engines publicly available for further research towards this new task.

As this evaluation can be seen as kick-off, there is obviously room for improvement. For example, in our study we restricted the dataset to researchers that already published in the field of the statement's topic. Further research needs to expand this to other people who can be experts that have not published papers in the fields of the statements' topics but deal with them such as journalists or politicians. In addition to that, we always measured a researcher's expertise by representing them as the content of exactly one publication (and its context). Naturally, future work needs to incorporate multiple publications of a researcher, e.g. to examine whether other similar works are sufficient to predict a researcher's expertise.

Acknowledgments

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the projects ReCAP and ReCAP-II, Grant Number 375342983 - 2018-2024, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

References

[1] M. Scharkow, F. Mangold, S. Stier, J. Breuer, How social network sites and other online intermediaries increase exposure to news, Proceedings of the National Academy of Sciences 117 (2020) 2761–2763.

[2] U. of Baltimore, University of baltimore: "fake news" has a real cost - and it's in the billions (22.11.2019), 2019. URL: https://www.ubalt.edu/news/news-releases.cfm?id=3425.

[3] J. Stachel, Einstein from'B'to'Z', volume 9, Springer Science & Business Media, 2001.

[4] J. Thorne, A. Vlachos, Automated fact checking: Task formulations, methods and future directions, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, Association for Computational Linguistics, 2018, pp. 3346–3359. URL: https://www.aclweb.org/anthology/C18-1283/.

[5] H. Ji, R. Grishman, Knowledge base population: Successful approaches and challenges, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, The Association for Computer Linguistics, 2011, pp. 1148–1158. URL: https://www.aclweb.org/anthology/P11-1115/.

[6] G. Angeli, C. D. Manning, Naturalli: Natural logic inference for common sense reasoning, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 534–545. URL: https://doi.org/10.3115/v1/d14-1059. doi:10.3115/v1/d14-1059.

[7] I. Dagan, B. Dolan, B. Magnini, D. Roth, Recognizing textual entailment: Rational, evaluation and approaches - erratum, Nat. Lang. Eng. 16 (2010) 105. URL: https://doi.org/10.1017/S1351324909990234. doi:10.1017/S1351324909990234.

[8] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak, V. Sable, C. Li, M. Tremayne, Claimbuster: The first-ever end-to-end fact-checking system, Proc. VLDB Endow. 10 (2017) 1945–1948. URL: http://www.vldb.org/pvldb/vol10/p1945-li.pdf. doi:10.14778/3137765.3137815.

[9] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2931–2937.

[10] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers, Association for Computational Linguistics, 2017, pp. 422–426. URL: https://doi.org/10.18653/v1/P17-2067. doi:10.18653/v1/P17-2067.

[11] Y. Long, Q. Lu, R. Xiang, M. Li, C. Huang, Fake news detection through multi-perspective speaker profiles, in: G. Kondrak, T. Watanabe (Eds.), Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers, Asian Federation of Natural Language Processing, 2017, pp. 252–256. URL: https://www.aclweb.org/anthology/I17-2043/.

[12] L. Zhou, J. K. Burgoon, J. F. Nunamaker, D. Twitchell, Automating linguistics-based cues

for detecting deception in text-based asynchronous computer-mediated communications, Group decision and negotiation 13 (2004) 81–106.

[13] M. Ott, Y. Choi, C. Cardie, J. T. Hancock, Finding deceptive opinion spam by any stretch of the imagination, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, The Association for Computer Linguistics, 2011, pp. 309–319. URL: https://www.aclweb.org/anthology/P11-1032/.

[14] R. Mihalcea, C. Strapparava, The lie detector: Explorations in the automatic recognition of deceptive language, in: ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers, The Association for Computer Linguistics, 2009, pp. 309–312. URL: https://www.aclweb.org/anthology/P09-2078/.

[15] M. Ott, C. Cardie, J. T. Hancock, Negative deceptive opinion spam, in: L. Vanderwende, H. D. III, K. Kirchhoff (Eds.), Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, The Association for Computational Linguistics, 2013, pp. 497–501. URL: https://www.aclweb.org/anthology/N13-1053/.

[16] S. Feng, R. Banerjee, Y. Choi, Syntactic stylometry for deception detection, in: The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers, The Association for Computer Linguistics, 2012, pp. 171–175. URL: https://www.aclweb.org/anthology/P12-2034/.

[17] Z. Gyöngyi, H. Garcia-Molina, J. O. Pedersen, Combating web spam with trustrank, in: M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, K. B. Schiefer (Eds.), (e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB 2004, Toronto, Canada, August 31 - September 3 2004, Morgan Kaufmann, 2004, pp. 576–587. URL: http://www.vldb.org/conf/2004/RS15P3.PDF. doi:10.1016/B978-012088469-8.50052-8.

[18] M. Cha, H. Haddadi, F. Benevenuto, K. Gummadi, Measuring user influence in twitter: The million follower fallacy, in: Proceedings of the international AAAI conference on web and social media, volume 4, 2010, pp. 10–17.

[19] K. R. Canini, B. Suh, P. Pirolli, Finding credible information sources in social networks based on content and social structure, in: PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011, IEEE Computer Society, 2011, pp. 1–8. URL: https://doi.org/10.1109/PASSAT/SocialCom.2011.91. doi:10.1109/PASSAT/SocialCom.2011.91.

[20] K. A. Ericsson, R. R. Hoffman, A. Kozbelt, A. M. Williams, The Cambridge handbook of expertise and expert performance, Cambridge University Press, 2018.

[21] E. J. Wright, K. B. Laskey, Credibility models for multi-source fusion, in: 9th International Conference on Information Fusion, FUSION 2006, Florence, Italy, July 10-13, 2006, IEEE, 2006, pp. 1–7. URL: https://doi.org/10.1109/ICIF.2006.301693. doi:10.1109/ICIF.2006.301693.

[22] C. Macdonald, I. Ounis, Voting for candidates: adapting data fusion techniques for an expert search task, in: P. S. Yu, V. J. Tsotras, E. A. Fox, B. Liu (Eds.), Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006, ACM, 2006, pp. 387–396. URL: https://doi.org/10.1145/1183614.1183671. doi:10.1145/1183614.1183671.

[23] D. Schall, F. Skopik, S. Dustdar, Expert discovery and interactions in mixed service-oriented systems, IEEE Trans. Serv. Comput. 5 (2012) 233–245. URL: https://doi.org/10.1109/TSC.2011.2. doi:10.1109/TSC.2011.2.

[24] L. Dumani, T. Wiesenfeldt, R. Schenkel, Fine and coarse granular argument classification before clustering, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, ACM, 2021, pp. 422–432. URL: https://doi.org/10.1145/3459637.3482431. doi:10.1145/3459637.3482431.

[25] M. I. Zegwaard, M. J. Aartsen, M. H. Grypdonck, P. Cuijpers, Differences in impact of long term caregiving for mentally ill older adults on the daily life of informal caregivers: a qualitative study, BMC psychiatry 13 (2013) 1–9.

[26] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977) 159–174.

[27] S. E. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.

[28] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[29] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/v1/D19-1410.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[31] D. Turnbull, Bm25f in lucene with blendedtermquery, 2016. URL: https://opensourceconnections.com/blog/2016/10/19/bm25f-in-lucene/.

[32] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[33] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, A. Jatowt, Yake! collection-independent automatic keyword extractor, in: G. Pasi, B. Piwowarski, L. Azzopardi, A. Hanbury (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2018, pp. 806–810.

[34] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! key-

word extraction from single documents using multiple local features, Information Sciences 509 (2020) 257–289. URL: https://www.sciencedirect.com/science/article/pii/S0020025519308588. doi:https://doi.org/10.1016/j.ins.2019.09.013.

[35] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, A. Jatowt, A text feature based automatic keyword extraction method for single documents, in: G. Pasi, B. Piwowarski, L. Azzopardi, A. Hanbury (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2018, pp. 684–691.