


Please cite the Published Version

Williams, Ashley  and Shardlow, Matthew (2022) Extending a corpus for assessing the credibility of software practitioner blog articles using meta-knowledge. In: EASE 2022: The International Conference on Evaluation and Assessment in Software Engineering, 13 June 2022 - 15 June 2022, Gothenburg, Sweden.

DOI: <https://doi.org/10.1145/3530019.3535310>

Publisher: Association for Computing Machinery (ACM)

Version: Accepted Version

Downloaded from: <https://e-space.mmu.ac.uk/629937/>

Usage rights:  In Copyright

Additional Information: © ACM 2022. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in EASE '22: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022, <http://dx.doi.org/10.1145/10.1145/3530019.3535310>.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Extending a corpus for assessing the credibility of software practitioner blog articles using meta-knowledge

Ashley Williams

ashley.williams@mmu.ac.uk

Manchester Metropolitan University
Manchester, UK

Matthew Shardlow

m.shardlow@mmu.ac.uk

Manchester Metropolitan University
Manchester, UK

ABSTRACT

Practitioner written grey literature, such as blog articles, has value in software engineering research. Such articles provide insight into practice that is often not visible to research. However, a high quantity and varying quality are two major challenges in utilising such material. Quality is defined as an aggregate of a document's relevance to the consumer and its credibility. Credibility is often assessed through a series of conceptual criteria that are specific to a particular user group. For researchers, previous work has found 'argumentation' and 'evidence' to be two important criteria.

In this paper, we extend a previously developed corpus by annotating at broader granularity. We then investigate whether the original annotations (sentence level) can infer these new annotations (article level). Our preliminary results show that sentence-level annotations infer the overall credibility of an article with an F1 score of 91%. These results indicate that the corpus can help future studies in detecting the credibility of practitioner written grey literature.

CCS CONCEPTS

• **General and reference** → **Cross-computing tools and techniques; Empirical studies**; • **Human-centered computing** → **Blogs**; • **Computing methodologies** → **Supervised learning**.

KEYWORDS

credibility assessment, argumentation mining, experience mining

1 INTRODUCTION

The Web has brought with it a shift in the way that software practitioners communicate. Storey et al. [17] describe that where practitioners once relied on face to face meetings and emails to disseminate their thoughts and opinions, they now utilise social media,

blog posts and online forums. This has led to an abundance of experience and practitioner insight being published on the Web. However, this rich data source is largely untouched by research.

Surveying such practitioner generated grey literature could provide new evidence and novel insight into practice. Garousi *et al.* [6] advocate the assessment of grey literature as a method for reducing publication bias and bridging the gap between the state-of-art, where research operates and the state-of-practice (i.e., what actually happens in industry). The same authors have provided guidelines for conducting Multi-Vocal Literature Reviews (MLRs) in software engineering research [7]. Surveying grey literature also provides instant, low-cost access to distant populations [21].

There are challenges in using grey literature in research however. Rainer and Williams [15] state the largest of these challenges to be identifying the quality content from the vast quantity available on the Web. There is therefore a need for systems to be developed that help us to identify and evaluate high-quality content and distinguish it from low-quality content. We define quality here in terms of a document's relevance to the consumer, and its credibility.

Credibility is a subjective concept. Credibility researchers have so far failed to agree on a definition of credibility, which has led to many candidate definitions which often conflict [18]. Research handles this subjectivity by instead assessing a series of conceptual criteria (e.g. bias, reasoning, citation) that are relevant to a particular user group (e.g. the visually impaired [1], first year students [10], pensioners [9]). The drawback of such an approach however is that the challenge is simply shifted to how we weight the criteria within that process, and then aggregate the criteria into an overall ranking so that we can compare the credibility of documents against each other. Further to this, credibility is subjective to the individual. Even within a single user group, credibility is subjective, and criteria may change over time. Two possible options are 1) to present objective criteria assessment in tables that allow the individual to choose how they combine the criteria into a overall rank (e.g. University and hospital ranking tables in the UK), and 2) automatically extracting meta-knowledge from the identified criteria.

We extend the corpus first introduced by Williams et al. [20] by reannotating the original texts at an article level. The new annotations arising from this are used to evaluate the definitions proposed in that work and propose strategies for composing article level meta-knowledge from sentence-level annotations. The paper investigates the following overarching research question:

RQ1 Can the credibility of practitioner written grey literature be determined through an information extraction approach, leveraging meta-knowledge?

The paper makes the following contributions:

- We extend a publicly available corpus of practitioner written blog articles with new annotations at an article level¹ so that we can look at meta-knowledge as an approach to ranking and aggregating credibility criteria.
- We present new evaluations on using the corpus to identify both individual labels, and overall document credibility.
- We present a review of articles annotated as not credible. The review provides suggestions and themes which are important for appearing credible in online articles.

2 RELATED WORK

2.1 Evidence in software engineering research

Drawing inspiration from the medical domain, Evidence Based Software Engineering (EBSE) integrates the best evidence from research with practical experience and human values [4]. Wohlin [22] presents an evidence profile for software engineering, though he acknowledges that synthesising available evidence is difficult, even in tightly controlled experiments. Wohlin’s evidence profile contains five considerations for evaluating evidence: Quality of evidence (is the evidence reliable?); Relevance of evidence; Ageing of evidence (is the evidence still relevant?); Vested interest/bias of the evidence provider; Strength of evidence.

Similarly, Fenton, Pfleeger and Glass [5] provide five questions to ask about any claim made in software engineering:

- (1) Is the claim based on empirical evaluation and data?
- (2) Was the empirical study designed correctly?
- (3) Is the claim based on a toy or real situation?
- (4) Were the measurements used appropriate to the goals of the empirical study?
- (5) Was the empirical study run for a long enough time?

Both lists provide insight into evidence requirements in research.

2.2 Practitioner written grey literature as evidence

Practitioners are often used as a source of evidence in research. Traditionally, such data is collected through methods such as surveys and interviews. However, the web has brought with it a shift in the way that practitioners communicate and disseminate their thoughts, opinions and discussions. This has led to the web being a rich source of data, and yet one which is often not utilised in research due to its varying quality (see Rainer and Williams [15] for a summary of benefits and challenges).

While researchers evidence their claims using facts and data, Devanbu *et al.* [3] and Rainer *et al.* [13] have observed that practitioners form opinions based on their personal and professional experiences. Practitioners may also be influenced by the experiences of their peers. This has led some researchers to question the value of grey literature as evidence. However, we argue that extracting practitioner opinion through survey and interview is similar to analysing grey literature.

Interest in grey literature has brought with it Grey Literature Reviews (GLRs) and Multi-vocal Literature Reviews (MLRs). GLRs concern systematically reviewing grey literature in order to understand the opinions of industry. MLRs seek to merge the views of

both research and grey literature. Garousi *et al.* [7] present guidelines for conducting MLRs in order to bridge the gap between research and industry. Unlike traditional systematic reviews, MLRs do not assume primary studies as their document of analysis and instead combine a mixture of grey literature documents and primary studies. This can be problematic as there is little to no quality control on grey literature documents, meaning care has to be taken when comparing their content to peer-reviewed primary studies. One solution could be to first conduct a GLR, or case survey, and then integrate the findings with the primary studies.

2.3 Assessing the quality of grey literature

We define ‘quality’ in terms of a documents relevance to the consumer (i.e. the researcher conducting a study using grey literature) and the documents credibility. For finding relevant content, we suggest a multi-faceted search approach using modern search engines [14] (though the term ‘relevance’ has multiple constructs in information science).

Assessing the credibility of a document is more challenging as it is a subjective concept and research has so far failed to agree on a definition. Further to this, a distinction is made between actual credibility, concerning fact checking the document (e.g. fake news detection) and perceived credibility, the internal steps a consumer takes towards assessing the document based on its source, content, method of delivery, and audience (e.g. social engagement) [11]. Research on credibility handles such subjectivity by reporting on (mostly) objective conceptual criteria (e.g. bias, argumentation, citations) that are relevant to a specific user group (e.g. the visually impaired [1], first year students [10], pensioners [9]). In the next section, we summarise results from a survey of software engineering researchers that aimed to determine the conceptual criteria important to them when assessing grey literature.

3 HOW DO SOFTWARE ENGINEERING RESEARCHERS ASSESS THE CREDIBILITY OF PRACTITIONER WRITTEN GREY LITERATURE?

Williams and Rainer [19] conducted a survey of software engineering researchers to investigate their opinions on the credibility of blog articles and identify the conceptual criteria that researchers adopt when evaluating the credibility of blog articles.

Candidate criteria were identified through a literature review that included thirteen papers selected for analysis. The literature review identified 88 conceptual criteria which were grouped into 9 categories for the survey. The survey was advertised to the programme committees of two international conferences on empirical software engineering and received 43 responses (44 response with 1 removed outlier). Participants were asked to rate the importance of each of the nine criteria.

The survey results show that researchers tend to place importance on the reasoning within an article and the evidence presented to support such reasoning. As expected, researchers place less importance on prior beliefs and the influence of others. The criteria from the survey were developed into a set of tags for annotating the original corpus [20].

¹<https://github.com/serenpa/Blog-Credibility-Corpus>

Articles were principally annotated for argumentation and evidence, which were deconstructed into specific sub-dimensions as described below. Dictionary definitions were used initially, and then refined throughout the annotation process to develop definitions specific for the context of the study.

The Argumentation annotations were broken down into the following sub-dimensions: Claim, Reasoning and Conclusion. Claim is defined as *A statement or assertion*. Claims may be supported by some reasoning or evidence, and may also be reflective of a personal opinion. Reasoning often appears close to the claim that is supported by the reasoning. Reasoning typically supports a claim with some form of logical justification or explanation. Conclusion is defined as *A judgement or decision reached by reasoning* and represents some finality.

The Evidence dimension was broken down into the following dimensions: Experience, Event, Citation, Code Snippet, Reference to table or image, Data/statistic and Other. Experience is defined as *References to a personal and/or professional experience which is provided as evidence to support a claim, or reasoning*. Event is a thing that happened. This may relate to specific time-bound instances. e.g. "Last summer, while attending a conference...". Direct active verbs are also used to imply the occurrence of an Event "The boy went to the shops". Citation, Code Snippet, Reference to table or image and Data/statistic all refer to similar occurrences which are described by the name of the label and reflect distinct sub-categories of Evidence. A final label called Other, allowed annotators to refer to other forms of evidence not captured by the labels.

4 CORPUS DEVELOPMENT

4.1 Corpus generation

We adopt the corpus first presented at the 2021 conference on Evaluation and Assessment in Software Engineering (EASE) by Williams et al. [20] and extend the work to article level annotations.

The corpus consists of articles from the blog of a single software practitioner, Joel Spolsky. Spolsky is the co-founder and former CEO of Stack Overflow. His software company was the creator of Trello before it was sold to Atlassian in 2017 for \$425 million. Spolsky's blog 'Joel on Software' is widely read and highly regarded by the practitioner community. The blog was mainly active from 2000 to 2012, but still publishes occasional articles today (the last article published at the time of writing was posted in January 2022). Spolsky's articles are a mix of: opinion pieces on topics such as software, management and start-ups; advertisements for new products and events; and short casual posts intended for fun, or to provide updates to his audience on recent activities.

Though there are clear threats to using models trained on the writings of a single practitioner, Spolsky's blog was chosen as it is an exemplar of the kind of content which could be useful as evidence in research which provides new insights into practice (e.g., Rainer [12] demonstrated the value in analysing practitioner-generated content using a single article from 'Joel on Software'). In terms of Berlo's model of communication [2], focusing on one practitioner allows us to assess content credibility, while controlling the source and medium.

Table 1: Agreement between annotators.

Agreement between all three annotators							
	Q1	Q2	Q3	Q4	Q5	Q6	Q7
All agree Y	162	152	31	22	209	6	166
All agree N	3	2	68	121	1	128	1
# all agree	165	154	99	143	210	134	167
# articles	234	234	234	234	234	234	234
% all agree	70.51	65.81	42.31	61.11	89.74	57.26	71.37
Agreement between annotators through voting (i.e. where at least two annotators agree).							
	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Y	215	213	83	69	231	31	222
N	19	21	151	165	3	203	12
Sum	234	234	234	234	234	234	234

The articles were initially crawled and downloaded using a publicly available web crawler². Each of the 1023 articles were then extracted from the HTML using Pattern, a Python library for text mining and pre-processing³. 234 of these articles were annotated at a sentence level giving a total corpus size of 19,996 sentences.

4.2 Article level annotations

Annotation was conducted at an article level. Three annotators were asked to read the 234 articles. The annotators were all students who were studying technology related degrees. After each article had been read, we asked them to answer seven yes/no questions and provide any additional comments if they felt their answers required clarification. The seven questions asked were:

- Q1 Does the document contain any reasoning?
- Q2 Does the document contain mentions of professional experience?
- Q3 Does the document contain mentions of personal experience?
- Q4 Does the document contain other forms of evidence to support their arguments?
- Q5 In general, is the article well written?
- Q6 In your opinion, is the article biased in any way?
- Q7 In your opinion, is the article credible?

We looked at article level agreement in two different ways (Table 1). The first, shows the percentage agreement when all three annotators agree on an answer for each of the seven questions. For the second, we adopted a voting system where annotations were treated as an agreement if two of the three annotators agreed.

5 ARTICLES ANNOTATED AS NOT CREDIBLE

The voted agreement on the article-level annotations (Table 1) shows 12 out of the 234 articles that were labelled as being not credible. In this section we look at the characteristics of the 12 articles, and combined with the annotators comments, investigate why these articles are perceived to be not credible.

Table 2 shows the article-level annotations for each of the 12 non-credible articles. The table shows that most articles are considered to

²https://github.com/serenpa/coast_crawl

³<https://github.com/clips/pattern>

Table 2: Annotations for the 12 articles labelled as not credible through voting. Q7 refers to the credibility question.

Article ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7
549	Y	Y	Y	N	N	Y	N
253	Y	N	Y	N	Y	Y	N
866	Y	Y	Y	N	Y	N	N
911	Y	Y	Y	N	Y	Y	N
919	Y	Y	N	N	Y	N	N
564	Y	N	Y	N	Y	Y	N
985	Y	Y	N	Y	Y	N	N
864	Y	Y	N	N	Y	Y	N
907	Y	Y	Y	N	N	Y	N
536	N	Y	N	N	N	N	N
743	N	Y	N	N	Y	Y	N
801	N	Y	N	N	Y	Y	N
Total	9	10	6	1	9	8	0

be well-written (Q5), and include reasoning and personal experience (Q1, Q3). However, 8 out of the 12 are also labelled as being biased (Q6) which, as expected, suggests that author bias is a contributing factor in the annotators assessment of credibility. This is reflected in the annotators comments. Two of the three annotators provided comments to justify decisions on the 12 articles (Table 3). Annotator 1 commented on 9 out of 12, and annotator 2 commented on 2 out of the 12. Their comments mostly focus on author bias (3 comments), a lack of justification (3 comments), and ambiguous writing (2 comments). One of the annotators also indicated that one article was labelled as not credible due to it being used as a recruitment stunt. There is not enough data to draw any meaningful conclusions, but such analysis in the future could provide an interesting contrast to the survey described in Section 3.

6 CORPUS EVALUATION

We have so far described the article level annotations that we added to the corpus. These overlap in scope with the original sentence level annotations, but are defined at different levels of granularity, giving rise to differences in the semantics of each set of annotations. In this section we will focus on the prediction of the article level annotations, based on both the article texts themselves and the potential effect that sentence level annotations may have on improving article level annotation prediction.

This framework makes use of the concept from information extraction of meta-knowledge [16], where knowledge dimensions at a higher level of granularity can be inferred from relevant sub-dimensions. Whereas this has previously been applied to structured events, in our context we see the sub-dimensions as those given at the sentence level and the higher dimensions as those given at the article levels. To help us understand the role of meta-knowledge in credibility prediction, we looked at predicting article level credibility from 3 different perspectives: (1) Can the text of the articles be used to predict the article level annotations? (2) Can the sentence annotations be used to predict article level annotations? (3) Do the sentence annotations improve the article level predictions when combined with text features.

To generate features for the texts, we employed a standard BERT-base model taken from the Huggingface library. The Huggingface

library provides pre-trained transformer models, and an API to run these over custom texts. We generated an embedding for each article by running the full article text through BERT and obtaining the embedding for the prediction token, which is a culmination of the embeddings for each token in the sentence. We did not fine-tune BERT at either the masked-language-modelling level, or prediction level due to the small size of our corpus. Instead, we chose to use a Decision Tree classifier (which is typically robust under small corpus sizes and class imbalances). We employed K-fold cross-validation, with K set to 3 and we report on the mean accuracy, class-weighted F1 score and area-under-curve (AUC) score. We used the voting method described in Table 1 as the labels for our classification task as this maximised the amount of data available in each class.

The first column of results in Table 4 shows our experiments using solely the BERT embeddings. This reflects the ability of the decision tree to distinguish between articles where the answer was Yes or No to each of our seven questions. Questions 1, 2, 5, 6 and 7 attained reasonable accuracy and F1 scores, in the 0.7 to 0.9 range. This indicates that the decision tree made the correct prediction in the majority of cases. The accuracy was lower for Q3 and Q4 however, where it is clear that the text features alone were not sufficient for predicting the answers to these questions.

The next column of results in Table 4 shows our experiments to predict article level annotations based on sentence level annotations. We used the proportion of sentences containing each feature in an article as a feature for the article level classification. This resulted in stronger predictive performance in most cases, as denoted by the presence of a ‘*’ in the table. The sentence level annotations cover dimensions of importance to the article level annotations and so it is to be expected that there is relevant information contained in these classifications that can be used to predict the article level dimensions. The original annotators read the entire articles in order to give the sentence annotations, and this shows that the BERT model is able to offer similar information to the prediction algorithm as a human who has read and categorised the sentences.

The final column of results in Table 4 shows our results of incorporating sentence level features with text-level features. We did this by providing the decision tree with all possible features. Although this creates a larger decision space, which can lead to worse classification accuracies, the decision tree selects the most useful features at each step in the algorithm and ignores less useful features, making it robust to the dimensionality expansion problem [8]. In Table 4, we see that most questions do not receive a benefit from integrating both feature sets, except for Question 4, where the scores all improve by 5-10%.

7 DISCUSSION AND CONCLUSIONS

This paper aims to investigate the RQ: *Can the credibility of practitioner written grey literature be determined through an information extraction approach, leveraging meta-knowledge?* To answer this, we added additional article-level annotations to an existing corpus, and then evaluated the degree to which we can predict these annotations using three approaches: 1. Can the text of the articles be used to predict the article-level annotations? 2. Can the existing

Table 3: Annotator comments for the 12 articles labelled as not credible through voting.

Article ID	Annotators Comments
549	We did not have proof or any justification backing what he has explained beyond his own personal experience. The interns were guided to follow his strides too so we can not count much on that; Biased (to C#) but still considers other options like Java and sometimes not clear.
253	If they mentioned disadvantages to a bias, is it still biased?
866	It was all about him and why his idea is better. There is need for further justification. There was also alot of ambiguity in his write up.
911	-
919	He could not give justifications through other sources for his reasons on recruiting a programmer.
564	The feedback were exaggerated and looks more like a personal vendetta.
985	Referral to construx.
864	-
907	It is more of the writer’s preference rather than general acceptable practice.
536	Too ambiguous and confusing.
743	It was not from a general point of view.
801	Recruitment stunt.

Table 4: Classification at the article level using Bert Embeddings, Sentence Dimensions and both combined. Best results are marked with a “*”

Question	Bert Embeddings			Sentence Dimensions			Combined		
	Accuracy	F1	AUC	Accuracy	F1	AUC	Accuracy	F1	AUC
Q1	0.842	* 0.856	* 0.574	* 0.863	0.851	0.521	0.842	0.855	0.552
Q2	0.795	0.812	* 0.534	* 0.842	* 0.839	0.485	0.786	0.797	0.506
Q3	0.585	0.574	0.509	* 0.650	* 0.636	* 0.597	0.607	0.577	0.544
Q4	0.585	0.561	0.492	0.581	0.555	0.421	* 0.637	* 0.594	* 0.532
Q5	0.957	0.959	0.485	* 0.970	* 0.970	* 0.489	0.962	0.961	0.478
Q6	0.714	0.730	0.426	* 0.774	* 0.771	* 0.510	0.748	0.722	0.439
Q7	0.893	0.900	0.497	* 0.919	* 0.909	* 0.519	0.897	0.907	0.508

sentence-level annotations be used to predict article-level annotations? and 3. Do the sentence annotations improve the article-level predictions when combined with text features.

We present three metrics to highlight the key outputs from our results. The accuracy and weighted F1 scores in Tables 4 demonstrate that we are able to make the correct prediction in the majority of cases, with some questions receiving higher scores than others. However, the AUC metric, which denotes the area under the ROC curve indicates that the classification performance suffered at times. This is probably due to the large class imbalance within the data, making it harder for the system to create reliable classifiers. We show that the sentence-level annotations are better predictors for the article-level annotations than text features. However, this may be affected by the bias introduced to the annotators by the question texts that are presented to them at annotation time. For example, we expect both sets of annotators to agree on articles that contain reasoning, experience and evidence. In answering questions 1–6 before assessing the credibility, annotators are already thinking about credibility in terms of the criteria discussed in previous questions.

Analysing the text without the sentence-level annotations appears to be a good predictor in some cases, but not always. This could be due to the class imbalance within the data as we find that the results are poorer for questions that have a more balanced representation in the corpus (e.g. for questions 3 and 4).

Interestingly, we see better results for question 4 when using a combination of the text and the sentence-level annotations. Question 4 asked the annotators to identify whether the article contained forms of evidence that supported argumentation. This is clearly a combination of other sentence level dimensions (argumentation and evidence) and so it makes sense that this would be well predicted by the sentence level dimensions. Further features from the text were also of use to the classifier here.

Our research has so far required hand-labelled dimensions for evaluation. This would clearly not be practical when scaling up to any real world application. Instead, the credibility criteria would need to be automatically predicted before being able to use them to predict overall article predictions. This would introduce more noise into the classification process. Future research will explore methods for assessing perceived credibility at article level.

Overall, the results are promising and provide clear direction for future research. However, more work is needed before such a system could be implemented for any practical application. Our next steps will focus on expanding the corpus and addressing threats as exploration continues.

7.1 Threats to validity

While the work presented in this paper serves as an important step towards our goals in credibility assessment, there are many threats in the works current form that need to be taken into consideration as we move forward:

Threats with the corpus There is a heavy class imbalance within the corpus which is affecting the results presented in our evaluation and models. This imbalance will be addressed as we scale up the size of the corpus to include multiple practitioners and domains. Further to this, the annotators used for both levels of annotation were students. Hence, they were not from our target demographic of software engineering researchers. This was not so much of a problem during the sentence level annotations as we provided annotators with strict criteria. However, asking them to provide a simple Yes/No for their assessment of the articles credibility is problematic because as we have discussed previously, students may have a different idea of what is credible to researchers. Again, this will be addressed as we scale up the corpus.

Inheriting threats from previous research The credibility criteria used for annotation throughout this research has stemmed from a literature review and survey of software engineering researchers [19]. In doing so, we inherit threats from that study. The two key threats that have greatest effect on this paper are 1) that the literature review was not conducted systematically. Therefore, we cannot be certain that there are not other credibility criteria missing from our analysis; 2) that the 43 survey respondents are not representative of the entire software engineering domain.

Single practitioner So far the corpus contains articles from a single practitioner. There are two reasons for this: firstly, because the chosen blog is an exemplar for showcasing our aims for the research; and secondly, because it focuses our analysis on the content of the article by controlling any influencing author or medium credibility. However, in only having a single practitioners articles there are obvious questions around how results may generalise, and also the bias in models built on the corpus (i.e. the corpus presents a single point of view from a single demographic).

7.2 Future research

A natural next step for this project it to scale the corpus with multiple practitioners. This would aid in addressing the issues of bias in models trained on the corpus. It would also allow us to investigate whether the results generalise to other practitioners who may have different writing styles/types of articles. In addition to this, we plan to extend the corpus to other domains so that we can make further generalisations on whether the results are specific to blog articles on software engineering.

In terms of specific applications for the corpus, we are interested in investigating the skills gap between academia and industry. Such a corpus could be useful in analysing the quality of grey literature prior to extracting industry trends, news and emerging technologies.

7.3 Conclusions

We present a corpus that allows us to determine the credibility of software practitioner grey literature. We also present new article-level annotations that build on an earlier version of the corpus and work towards investigating whether article credibility can be inferred from sentence level annotations. Our results show a high accuracy and F1 score. However, the area under the ROC curve is low. This is due to the large class imbalance within the data. Future work will address the class imbalance and other threats by

scaling the size of the corpus up to include multiple practitioners and domains.

REFERENCES

- [1] Ali Abdolrahmani and Ravi Kuber. 2016. Should I trust it when I cannot see it? Credibility assessment for blind web users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 191–199.
- [2] David K Berlo. 1960. *The Process of Communication: An Introduction to Theory and Practice*. Rinehart Press, San Francisco, USA.
- [3] Premkumar Devanbu, Thomas Zimmermann, and Christian Bird. 2016. Belief & evidence in empirical software engineering. In *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*. IEEE, ACM, Limerick, Ireland, 108–119.
- [4] Tore Dyba, Barbara A Kitchenham, and Magne Jorgensen. 2005. Evidence-based software engineering for practitioners. *IEEE software* 22, 1 (2005), 58–65.
- [5] Norman Fenton, Shari Lawrence Pfleeger, and Robert L. Glass. 1994. Science and substance: A challenge to software engineers. *IEEE software* 11, 4 (1994), 86–95.
- [6] Vahid Garousi, Michael Felderer, and Mika V Mäntylä. 2016. The need for multivocal literature reviews in software engineering: complementing systematic literature reviews with grey literature. In *Proceedings of the 20th international conference on evaluation and assessment in software engineering*. ACM, Limerick, Ireland, 1–6.
- [7] Vahid Garousi, Michael Felderer, and Mika V Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* 106 (2019), 101–121.
- [8] Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, Dallas, USA, 604–613.
- [9] Qingzi Vera Liao. 2010. Effects of cognitive aging on credibility assessment of online health information. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 4321–4326.
- [10] Ericka Menchen-Trevino and Eszter Hargittai. 2011. YOUNG ADULTS' CREDIBILITY ASSESSMENT OF WIKIPEDIA. *Information, Communication & Society* 14, 1 (2011), 24–51.
- [11] Miriam J Metzger. 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American society for information science and technology* 58, 13 (2007), 2078–2091.
- [12] Austen Rainer. 2017. Using argumentation theory to analyse software practitioners' defeasible evidence, inference and belief. *Information and Software Technology* 87 (2017), 62–80.
- [13] Austen Rainer, Tracy Hall, and Nathan Baddoo. 2003. Persuading developers to "buy into" software process improvement: a local opinion and empirical evidence. In *2003 International Symposium on Empirical Software Engineering, 2003. ISESE 2003. Proceedings*. IEEE, IEEE, Rome, Italy, 326–335.
- [14] Austen Rainer and Ashley Williams. 2019. Heuristics for improving the rigour and relevance of grey literature searches for software engineering research. *Information and Software Technology* 106 (2019), 231–233.
- [15] Austen Rainer and Ashley Williams. 2019. Using blog-like documents to investigate software practice: Benefits, challenges, and research directions. *Journal of Software: Evolution and Process* 31, 11 (2019), e2197.
- [16] Matthew Shardlow, Riza Batista-Navarro, Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2018. Identification of research hypotheses and new knowledge from scientific literature. *BMC medical informatics and decision making* 18, 1 (2018), 1–13.
- [17] Margaret-Anne Storey, Leif Singer, Brendan Cleary, Fernando Figueira Filho, and Alexey Zagalsky. 2014. The (r) evolution of social media in software engineering. In *Future of Software Engineering Proceedings*. ACM, New York, NY, USA, 100–116.
- [18] Ashley Williams and Austen Rainer. 2017. *The analysis and synthesis of previous work on credibility assessment in online media: technical report*. Technical Report. Technical Report. University of Canterbury, NZ.
- [19] Ashley Williams and Austen Rainer. 2019. How do empirical software engineering researchers assess the credibility of practitioner-generated blog posts? In *Proceedings of the Evaluation and Assessment on Software Engineering*. ACM, Copenhagen, Denmark, 211–220.
- [20] Ashley Williams, Matthew Shardlow, and Austen Rainer. 2021. Towards a corpus for credibility assessment in software practitioner blog articles. In *Evaluation and Assessment in Software Engineering*. ACM, Trondheim, Norway, 100–108.
- [21] Elena Wilson, Amanda Kenny, and Virginia Dickson-Swift. 2015. Using blogs as a qualitative health research tool: a scoping review. *International journal of qualitative methods* 14, 5 (2015), 1609406915618049.
- [22] Claes Wohlin. 2013. An evidence profile for software engineering research and practice. In *Perspectives on the Future of Software Engineering*. Springer, Berlin, Heidelberg, 145–157.