Towards a corpus for credibility assessment in software practitioner blog articles

Ashley Williams
ashley.williams@mmu.ac.uk
Department of Computing and Mathematics
Manchester Metropolitan University
UK

Matthew Shardlow
m.shardlow@mmu.ac.uk
Department of Computing and Mathematics
Manchester Metropolitan University
UK

Austen Rainer
a.rainer@qub.ac.uk
School of Electronics, Electrical Engineering and Computer Science
Queens University Belfast
UK

ABSTRACT

Background: Blogs are a source of grey literature which are widely adopted by software practitioners for disseminating opinion and experience. Analysing such articles can provide useful insights into the state–of–practice for software engineering research. However, there are challenges in identifying higher quality content from the large quantity of articles available. Credibility assessment can help in identifying quality content, though there is a lack of existing corpora. Credibility is typically measured through a series of conceptual criteria, with 'argumentation' and 'evidence' being two important criteria.

Objective: We create a corpus labelled for argumentation and evidence that can aid the credibility community. The corpus consists of articles from the blog of a single software practitioner and is publicly available.

Method: Three annotators label the corpus with a series of conceptual credibility criteria, reaching an agreement of 0.82 (Fleiss' Kappa). We present preliminary analysis of the corpus by using it to investigate the identification of claim sentences (one of our ten labels).

Results: We train four systems (Bert, KNN, Decision Tree and SVM) using three feature sets (Bag of Words, Topic Modelling and InferSent), achieving an F1 score of 0.64 using InferSent and a Linear SVM.

Conclusions: Our preliminary results are promising, indicating that the corpus can help future studies in detecting the credibility of grey literature. Future research will investigate the degree to which the sentence level annotations can infer the credibility of the overall document.

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, text mining

1 INTRODUCTION

1.1 Context and motivation

Garousi *et al.* [15] argue that the wider adoption of grey literature by research would help to bridge the gap between the state–of–art, where research typically operates, and the state–of–practice, what actually happens in industry. Conducting grey literature reviews systematically is problematic, however, as we have limited visibility into how search engines rank results, and the web is a lot larger and more diverse than academic databases. Rainer and Williams [30] suggest that the use of credibility assessment as an inclusion criterion in such grey literature reviews can contribute to providing a method for identifying the higher quality results.

Credibility is, at least in part, a subjective experience. Credibility research attempts to address this subjectivity by reporting credibility assessment concerning a specific user group e.g. the visually impaired [1], first year students [24], pensioners [20]. However, conceptualising credibility at the level of a user group is itself problematic because members of the user group still have unique experiences in time and space. The subjectivity of credibility affects our ability to assess the credibility of web documents, e.g., our ability to consistently annotate web documents. We return to this point in Section 2.3.

Credibility is assessed through measuring a series of conceptual criteria. The specific criteria to use depends on the user group and the context of the study being undertaken. Williams and Rainer [43] conducted a survey in order to determine which criteria apply to the credibility assessment of software engineering researchers. Both reasoning and the reporting of experience were considered as important criteria. For measuring reasoning, we can turn to the argumentation mining community for prior research. Measuring experience is more challenging as, according to Rainer *et al.* [31], there is limited prior research on experience mining.

In this paper, we present a corpus of annotated articles from a single software practitioner's blog (the highly regarded 'Joel On

Software' blog, written by Joel Spolsky). The articles have been annotated at a sentence level for the presence of argumentation and evidence. The corpus contains 19996 sentences and is publicly available for download. We present results from our preliminary use of the corpus, discuss how the corpus and results build on existing credibility assessment research, contrast the annotations with the results of an existing tool (MARGOT [22]), and consider next steps.

1.2 Contributions

The paper makes the following contributions:

- We present a new corpus of 19996 sentences, annotated for argumentation and evidence. The corpus comprises all sentences from 234 blog articles (the full blog roll comprises over 1000 blog articles). The corpus is available for public use[1].
- To the best of our knowledge, the annotated corpus is the first corpus of blog documents written by an experienced software practitioner and annotated for arguments and evidence.
- We also present our preliminary work towards credibility assessment by identifying claim sentences within the corpus.

1.3 Structure of this paper

The remainder of this paper is structured as follows: section 2 provides an overview of previous and related work; section 3 describes the generation of a novel annotated dataset; section 4 describes the design of our study; section 5 presents and discusses the results of our study; the conclusions and future work are presented in section 6.

2 RELATED WORK

2.1 Industry–originating research fields

Grey literature that has been written by experienced software practitioners often influences industry practice. In some cases, ideas originating in industry create and mould new research directions. For example, in 2014, Martin Fowler, a prolific software practitioner, published a blog article on the microservices' architecture for scalable web applications [19]. According to the blog article, the term "microservices" had been discussed and agreed on by a group of software architects to describe a common architectural style that many of them had been recently exploring. Since the articles' release, microservices have not only become a powerful and widely used architectural pattern in industry (e.g. Netflix famously uses microservices[2], as well as Monzo[3]), but also microservices have acquired a large research community (e.g. [13, 35, 44]). Microservices are one example of practice influencing and moulding research.

2.2 The use of practitioner–generated grey literature in research

Software engineering researchers often use practitioners as a source of evidence in their studies. Typically, such evidence is collected through traditional evidence gathering techniques, e.g. survey, interview, observation. The world wide web has brought with it a shift in the way that software practitioners disseminate information [36]. As a result, there is growing interest in utilising grey literature as an additional data source in evidence gathering through Grey Literature Reviews (GLR; e.g., [2, 3, 6, 35]) and Multi–vocal Literature Reviews (MLR; e.g., [16]).

There are however challenges in looking at grey literature. Rainer and Williams [30] describe the challenges of looking at blog–like content. These challenges generalise to all grey literature and are summarised here:

- Definitions and models – there exists multiple, and sometimes conflicting definitions for 'grey literature.' There are also a lack of models to describe grey literature structure and relationships.
- Classification frameworks – as well as the definitions, there are discrepancies in the literature over how the quality of different grey literature sources compare. Garousi *et al.* [16] present a framework for classifying grey literature which builds on existing frameworks [2, 3]. The frameworks imply a hierarchy of quality within different grey literature sources, but Rainer and Williams [30] argue that the quality of grey literature cannot be classified solely by the medium in which it is published.
- The quantity and quality of grey literature – the universe of grey literature is substantially larger than academic literature and its quality varies greatly. We need a reliable and rigorous way of filtering the high quality content from the vast quantity available.
- Ambiguity of language – grey literature is typically informal and therefore can use idiomatic structures, which may introduce ambiguity.

2.3 Credibility assessment

Assessing the credibility of a document can help distinguish the higher quality grey literature from the vast quantity available. Credibility assessment is subjective to the individual. The literature handles this subjectivity by reporting and assessing conceptual criteria for a particular user group (e.g. visually impaired, first year students, pensioners). Williams and Rainer [43] surveyed software engineering researchers to determine the most important conceptual criteria when assessing blog articles. Two key criteria identified were 1) the presence of the argumentation within the document, and 2) the evidence and personal experience provided to support the argumentation. Personal experience is important in grey literature because where researchers argue based on data and experiment, practitioners form opinions based on their personal and professional experience [11, 29]. For identifying arguments and experience within text, we can utilise the argumentation mining, opinion mining [5, 33] and experience mining communities. Lippi and Torroni [21] present a review of the state of the argumentation mining community. Lippi and Torroni also released MARGOT [22], a publicly available tool for assessing a document's argumentation and evidence. MARGOT was trained on the IBM Debater dataset, the largest corpus available at the time. The experience mining community is not as mature as its argumentation mining counterpart.

[1] https://github.com/serenpa/Blog-Credibility-Corpus
[2] https://netflixtechblog.com/tagged/microservices
[3] https://www.infoq.com/presentations/monzo-microservices/

Rainer *et al.* [31] conducted a review of the literature reporting the identification of professional experience in grey literature. The review concluded that more primary studies are needed in order to advance the community. One barrier to such primary studies is a lack of corpora that has been labelled for experience. This paper contributes in that it presents a new dataset which is publicly available and is annotated for argumentation and evidence.

2.4 Aggregating conceptual criteria measurements into overall credibility

The subjective nature of credibility also hinders the ability to aggregate conceptual criteria measurements into a score for overall credibility. Previous attempts at ranking using various techniques are easily criticised due to investigators deciding on the weightings of each criteria (e.g. [40]). One solution explored has been to measure and present conceptual criteria (e.g. reasoning and experience) data back to the researcher in tabular format with the ability to rank as they please (such as University ranking tables). Our future research intends to look towards other techniques, such as meta–knowledge [34], for ranking and comparing grey literature documents.

3 DATASET

3.1 The subject of the corpus

The corpus consists of articles from a single practitioner's blog, Joel Spolsky[4]. His 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 articles sporadically today (the last article published at the time of writing was June 2020). The articles within the blog are a mix of opinion pieces (on subjects such as software and technology, management, and start–ups), advertisements for new products and events, and short casual posts intended for fun, or to provide brief updates to his audience on his thoughts/recent activities. This mix of article types brings with it additional challenges over looking at the blogs of practitioners that maintain a more uniform structure (e.g. Martin Fowler[5]).

The blog was chosen due to previous research finding it to be an exemplar of how practitioner–generated content can provide new insights for research. Rainer [28] demonstrated the value in analysing practitioner–generated content using a single article from 'Joel on Software,' and Williams [39] used the blog to evaluate a set of keywords for identifying reasoning.

3.2 Data gathering

The data was initially collected using `COAST_CRAWL`[6], a publicly available web crawler. We seed the crawler with the blog's archive page to ensure that all articles are accessible. The crawler initially finds 1693 pages. However, after de–duplication, removing static pages and removing non–article pages, 1023 remain. The article text is extracted using Pattern[7].

[4]Spolsky is the co–founder and former CEO of Stack Overflow, and the co–founder of Fog Creek Software (the company that created Trello)
[5]https://martinfowler.com/
[6]https://github.com/serenpa/coast_crawl
[7]https://github.com/clips/pattern

3.3 Annotation

3.3.1 Tagset. The articles were annotated for argumentation and evidence. These two criteria were broken down into more specific tags. Dictionary definitions for each of the tags were developed further throughout phase one of the annotations (Section 3.3.2). Table 1 lists each of the tags with a summary definition for each.

3.3.2 Process. Annotators were presented with the entire document. They were asked to read the article in its entirety before labelling the sentences which contain argumentation and evidence. WebAnno[8] [14] was used for annotation. Annotation was conducted in two phases. In phase one, two annotators were employed with an additional third annotator for resolving conflicts. This phase consisted of four rounds of annotation with each of the three annotators completing all of the articles from the round. At the end of each round, we met with the three annotators to discuss annotations as a group. Annotators were encouraged to converge on their annotations, but we allowed discrepancies between the two main annotators which were later resolved by the third annotator. We maintained a set of annotation guidelines which were updated during these meetings as the annotators' definitions evolved and converged. In total, 36 articles were annotated during this initial phase. At the end of the phase, agreement was calculated on the annotations at a sentence level. We used the third annotator to resolve the conflicts between annotators one and two. In instances where a conflict cannot be resolved (e.g., if annotator 3 did not agree with either label), we favour annotator one as they annotated the most articles overall. The agreement at the end of phase one for argumentation labels is 0.817 using Fleiss' Kappa (Po: 0.895; Pe: 0.428), and 0.819 for evidence labels (Po: 0.959; Pe: 0.771).

Phase two then consisted of the annotators working on different articles with no double annotation taking place. In doing so, a further 198 articles were annotated taking the total number of articles up to 234. This leads to a final dataset size of 19996 sentences.

3.3.3 Description of the dataset. Table 2 presents frequency counts and percentages for each of the labels. The table indicates that 34% of the sentences have no label. The percentages total more than 100% because approximately 1200 sentences are labelled with more than one label.

Table 3 presents a contingency table of sentences that are labelled with more than one label. Of the 19996 sentences, 1242 sentences have more than one label. We allowed annotators to give more than one label per sentence at their discretion as complex sentences may contain multiple elements to be tagged and therefore do not always align with sentence level annotations.

3.4 Comparing the annotated dataset with an independent system

As an exploration of the annotated dataset, we compared the labels assigned by the annotators with labels assigned independently by MARGOT [22], an established argumentation mining tool. (To clarify: we are not *evaluating* the two set of labels. We view this merely as a comparison to a related system from previous literature). We first used MARGOT to label the sentences of all articles in the Spolsky dataset. MARGOT actually 'generated' slightly more sentences

[8]https://webanno.github.io/webanno/

Table 1: The annotation tags and definitions

Tag (acronym)	Definition
Argumentation	
Claim (clm)	A statement or assertion. This claim may be supported by some reasoning or evidence, and may also be an opinion
Reasoning (rsn)	Reasons often (but not always) appear close to the claim that the reason is supporting. Reasoning supports a claim with logical justification/explanation
Conclusion (conc)	A judgement or decision reached by reasoning
Evidence	
Experience (exp)	References to a personal and/or professional experience which is provided as evidence to support a claim, or reasoning. We are interested here in actual experience (c.f. implied experience or hypothetical experience).
Event (evnt)	Events are defined as things that have happened. Operationally, we may detect events through specific mentions to a moment in time e.g. "Last summer, while attending a conference...". Verbs can also imply that an event has taken place without referencing a specific time e.g. "The boy went to the shops".
Citation (cite)	May be a URL hyperlink to a other web page, a formal reference in a dedicated references section typically at the end of an article (as in research), a footnote, an in–text citation (without a dedicated references section).
Code Snippet (code)	Authors may evidence their claims through the use of code examples. These code examples may be in–text, in a separate block, in an image, or in a table.
Reference to table or image (ref)	Authors may evidence their claims through the use of tables of data and/or images.
Data/statistic (data)	Authors may evidence their claims through providing statistics or presenting analysis or other forms of raw/processed data.
Other (othr)	There may be other forms of evidence which has not been specified in the guidelines. This tag allows for annotators to flag other forms of evidence they may think is relevant for discussion.

Table 2: Frequency counts and percentages for the annotation labels

Label	Frequency	Percentage
Claim	9202	62
Reasoning	1586	11
Conclusion	331	2
Citation	778	5
Code Snippet	61	0
Events	261	2
Experience	2590	17
Reference to Table or Image	29	0
Statistics or Data	22	0
Other	29	0
Total labelled	14889	
No label	5107	34
# documents annotated	234	-
# sentences with multiple labels	1242	-

Table 3: Contingency table (n=1242). The values of zero (0) in the table are included for completeness.

Label	Rsn	Conc	Cite	Code	Evnt	Exp	Ref	Data	Othr
Claim	643	61	168	44	57	119	2	3	19
Reasoning		109	19	9	9	53	0	0	2
Conclusion			3	1	0	3	0	0	0
Citation				2	8	88	0	0	3
Code Snippet					2	2	0	1	0
Events						20	0	1	0
Experience							0	2	5
Reference to Table / Image								0	0
Statistics / Data									0

(20022 compared to the 19996 sentences of the annotation dataset). We then selected the 234 articles that had been annotated by both the annotators and MARGOT. We matched similar sentences using Jaccard similarity: two sentences were treated as sufficiently similar if the Jaccard similarity score was >= 0.5. Our matching identified approximately 20000 matched pairs of sentences. For the matched sentences, we compared the labels assigned by the annotators with the labels assigned by MARGOT.

Table 4 presents confusion matrices for three sets of labels. In Table 4(a) we compare MARGOT's labelling of a claim (TRUE or FALSE) with the annotators' labelling of *either* a claim, a reason or a conclusion (cf. Table 1). In Table 4(b) we compare MARGOT's labelling of a claim with the annotators' labelling *only* of a claim. In Table 4(c) we compare MARGOT's labelling of a claim with the annotators' labelling *only* of either a reason or a conclusion.

Table 4 shows considerable disagreement between the annotators' labels and MARGOT. We hypothesise that this disagreement can at least partly be explained by definitions and by the nature of the dataset.

For definitions, we hypothesise that our definition of claim (cf. Table 1) is different to the definition applied during the labelling of the IBM Debater dataset that was subsequently used to train MARGOT. For the IBM Debater dataset, Aharoni *et al.* [4] defined a claim as: "Context Dependent Claim – a general concise statement that directly supports or contests the topic". This contrasts with our definition of a claim as, "A statement or assertion. This claim may be supported by some reasoning or evidence, and may also be an opinion." (see Table 1).

For the nature of the dataset, we hypothesise that the content of the Spolsky dataset is different to that of the IBM Debater dataset. Lippi and Torroni [22] explain that the IBM Debater dataset consisted of 547 Wikipedia articles that had been organized into 58 topics, and annotated with 2294 claims and 4690 evidence facts. By contrast, we have 234 web articles written by one software practitioner on an unknown number of topics. It is likely that the Wikipedia articles will have progressed through more review and revision than the web articles, and will also likely be written in a more formal style. We therefore hypothesise that the annotators' labels in Table 4(c) are most likely to be comparable to the MARGOT claims, however the respective data is too imbalanced to explore this hypothesis further at this stage.

4 PRELIMINARY EXPERIMENT

Our experiments focus on the detection of claims in our dataset. This is only one use of the dataset and the detection of other categories is left to future work (however, preliminary results for our other labels are presented in Table 6). The detection of claims is an interesting problem in itself as it is a broadly defined concept and many types of sentences can be considered a claim. In our corpus, 62% of sentences are annotated as claims. We use 4 approaches as detailed below:

BERT: The BERT masked language model [12], built on the transformer architecture [37] is now prevalent at the forefront of NLP research. We used the keras-bert implementation [9] with the Bert-Large pretrained model (L=24, H=1024, A=16) and configured it with one fully connected hidden layer of 16 units with a ReLU activation function and an output layer of 2 hidden units with a softmax activation function. We used the rectified Adam optimiser [23] during training. We report on the model with and without the hidden layer to demonstrate its effect.

K-Nearest Neighbour (KNN): We used the SciKit Learn [26] implementation of the KNN [17], with K set to 3.

Decision Tree (DT): We used the SciKit Learn implementation of the decision tree [8], with a maximum depth of 5.

Support Vector Machine (SVM): We used the SciKit Learn implementation of the SVM [9, 27] with a linear kernel and regularisation parameter $C = 0.25$.

We split the data into train (80%), validation (10%) and test (10%) partitions, which were stratified according to the claim label and shared across all algorithms. We removed any sentences with multiple labels, leaving a total of 8539 Claim sentences and 9612 sentences with no label (split evenly across the three partitions). The validation partition was used to measure the loss whilst training BERT and to select appropriate algorithms from Sci-kit learn. The final results are reported on the test partition. We trained Bert for one epoch in each case and did not further tune the hyperparameters of BERT or the other algorithms. Although this can lead to a perceived improvement in results, it also often leads to model overfitting, which we sought to avoid.

We created three feature sets that were used as input to our machine learning algorithms (excluding BERT, which does not require external features). These are described as follows:

Bag of Words: In this approach we first computed the mutual information [25] between each word and the class label. We selected the top 100 words as binary features, which indicated the presence of a word that distinguished the class.

Topic Modelling: We first created a document-token matrix indicating the frequency of each token in each sentence. We then used the gensim [32] implementation of LDA [7] to reduce the dimensionality of the matrix and to create topic vectors for each document in a technique commonly known as topic modelling. This technique forces words which occur in similar contexts to be in the same topic. We limited the vector size to 100 topics.

InferSent: We used the Facebook library InferSent [10], which provides a 4096 dimensional embedding for a given sentence. InferSent uses FastText vectors [18] to get the embedding for each token and then passes these through a pre–trained model which identifies the importance and weighting of each embedding before recombination. Each dimension of the resulting embedding was used as a feature to the algorithm, giving 4096 distinct features.

We ran each algorithm with each feature set and report the results in Table 5. We attempted to combine feature sets, but this did not lead to any improvement in the overall scores and so these results are omitted.

5 RESULTS & DISCUSSION

Our results are shown in Table 5. We tried using Bert as it is has been shown to achieve state–of–the–art performance with little fine

[9] https://github.com/CyberZHG/keras-bert

Table 4: Confusion matrices comparing the annotators' labels with MARGOT's [22] labels

(a) Contrasting MARGOT's label of a claim with annotators' labels of a claim, reason or conclusion.

	MARGOT		
Annotators	**TRUE**	**FALSE**	**Total**
TRUE	952	9344	10296
FALSE	297	9429	9726
Total	1249	18773	20022

(b) Contrasting MARGOT's label of a claim only with annotators' label of a claim

	MARGOT		
Annotators	**TRUE**	**FALSE**	**Total**
TRUE	825	8327	9152
FALSE	424	10446	10870
Total	1249	18773	20022

(c) Contrasting MARGOT's label of a claim only with annotators' label of a reason or conclusion

	MARGOT		
Annotators	**TRUE**	**FALSE**	**Total**
TRUE	255	1542	1796
FALSE	994	17220	18214
Total	1249	18773	20022

Classifier	Features	Precision	Recall	F1
BERT	No Hidden Layer	0.54	0.48	0.51
	Hidden Layer	0.75	0.48	0.58
KNN	BOW	0.88	0.12	0.20
	Topic Modelling	0.52	0.45	0.48
	InferSent	0.57	0.58	0.58
DT	BOW	0.82	0.02	0.04
	Topic Modelling	0.56	0.31	0.40
	InferSent	0.57	0.66	0.61
SVM	BOW	0.85	0.01	0.02
	Topic Modelling	0.00	0.00	0.00
	InferSent	0.62	0.65	**0.64**

Table 5: The results of detecting claim sentences

tuning for other NLP tasks. The results however, were somewhat disappointing. Even the addition of a hidden layer did not greatly improve the scores. Although we could have spent time further modifying the network structure and training for many epochs, we instead decided to use classic machine learning algorithms from sci–kit learn.

We used 2 traditional feature generation techniques (Bag of Words and Topic Modelling) and one state of the art method of generating sentence embeddings (InferSent). The InferSent features outperformed Bag of Words and Topic Modelling with every classifier. This is surprising as Bag of Words and Topic Modelling have 100 features each, compared to 4096 features for InferSent. Typically, the performance of classical machine learning algorithms decreases when presented with many features. We hypothesise that many of the dimensions in the InferSent embeddings were not being used as part of the classification strategy. Our best performing system used the InferSent embeddings and a Linear SVM and received an F1 score of 0.64, with precision at 0.62 and recall at 0.65. This indicates that the claim vs. non-claim sentences are separable and that a model can be built to distinguish between them. InferSent provides an embedding for a sentence and it would be interesting to analyse which parts of a sentence are being used in the classification of claims vs. non–claims.

In one instance, our model was not able to produce a reliable model for the Claim sub–dataset (see SVM + Topic Modelling). We used a linear SVM and this implies that no linear boundary could be found in the topic space to separate our classes. Note that the KNN and Decision Tree, both of which can create complex boundaries in feature space, were able to produce models using the topic model features. It may be the case that using a kernel–based SVM would yield a more reliable model for the Topic–Modelling

features, however we avoided non–linear SVM as the time taken to train is too great with our high–dimensional InferSent embeddings.

It may well be possible to improve our scores by tuning the algorithms used or using a more powerful Transformer based architecture, however we have focused our results on building a model for claims as a benchmark for our dataset. We expect to further improve on the scores we have reported as well as identify other elements in our corpus in future work.

More generally, credibility assessment is a difficult task as it is subjective, multi–disciplinary, and lacking in formal definitions. The credibility literature acknowledges that argumentation and the reporting of evidence are important criteria. However, when we try to apply them to this particular application, we find a relatively low amount of reasoning and evidence within the dataset. Furthermore, most studies within the argumentation mining community focus on well–structured text such as legal texts, persuasive essays and debate corpora. When applying technologies such as MARGOT, to web documents, the classification task becomes more difficult as the writing style is more casual, less structured and at least sometimes more ambiguous than one might expect of better structured texts.

Finally, the credibility assessment community recognises a need for formal definitions for both the conceptual criteria analysed and the term 'credibility' itself [41]. We see similar issues when contrasting our work with the argumentation mining community. For example, the IBM Debater dataset uses the term 'claim' as its measure of argumentation. In this paper, we define claim as synonymous with assertion and opinion, distinguishing it from reasoning and conclusion.

6 CONCLUSIONS & FUTURE RESEARCH

6.1 Threats to Validity

There are multiple threats to this research, with each threat also providing an opportunity for future work:

(1) The metrics used for assessing credibility are based on the findings of a previous literature review and survey [43]. The literature review was not conducted systematically and only thirteen papers were selected for analysis. A broader, systematic review may yield new important metrics for credibility assessment. In addition, although the response rate of the survey was good, the overall number of responses in comparison to the community of software engineering researchers is relatively low. Further research is needed to ensure that the credibility metrics do not only apply to this subset of researchers.
(2) There are threats with the way in which we have conducted our annotations. Our student annotators may have different definitions for our credibility metrics than software engineering researchers (our target demographic). This may in turn lead to different annotations. Future research will look at the quality of our annotations.
(3) In looking at only one source of grey literature, a single practitioners blog, it is unclear how well our models, and models trained using our dataset, generalise to other sources. Future research will investigate the degree to which our metrics and models work over other data sources (e.g. Twitter, Stack Overflow, GitHub).

6.2 Future research

There are many avenues open for future research. This paper presents preliminary analysis of one label within the dataset, further analysis across all labels is a natural next step. The analysis could then be aggregated and ranked to form an overall credibility rating for each individual article. The credibility ratings and quality could then be compared against one another.

We also plan to benchmark against other tools and to look at particular subsets of the dataset. For example, we plan to look more closely at sentences where MARGOT and the annotators agree. This is challenging as the result would be a small, imbalanced dataset, however it could provide further insight into identifying quality content.

Finally, the dataset presented in this paper can also aid further work in each sub–community (e.g. argumentation, experience, opinion mining). For example, previous work [38, 42] has investigated the degree to which practitioners cite research in their blog articles. The dataset allows for further, more in–depth analysis of citations. Similarly, another area for future research is the dataset's potential to be used as a source in future experience mining primary studies.

6.3 Conclusions

In this paper we have presented a new dataset annotated for elements of argumentation and evidence. The dataset comprises 19996 labelled sentences from 234 complete blog articles, with all articles written by an experienced software practitioner. Our intention is that the dataset can help future studies in automating credibility assessment and the comparison of documents for ranking based on credibility and quality. The dataset is publicly available[10].

In addition to the dataset generation, we present preliminary analysis toward automating the identification of claim sentences, one of our ten labels. An SVM trained using the InferSent feature set provides a F1 score of 0.64.

ACKNOWLEDGMENTS

We thank the annotators who contributed to this project, the Centre for Advanced Computational Science for their seed funding, and the reviewers for their valuable feedback.

We have sought advice on the ethics of analysing blogs. All of the blog articles that we have used are publicly available and we would like to thank Joel Spolsky, who is aware of this research.

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] Jean Adams, Frances C Hillier-Brown, Helen J Moore, Amelia A Lake, Vera Araujo-Soares, Martin White, and Carolyn Summerbell. 2016. Searching and synthesising 'grey literature'and 'grey information'in public health: critical reflections on three case studies. *Systematic reviews* 5, 1 (2016), 164.
[3] Richard J Adams, Palie Smart, and Anne Sigismund Huff. 2017. Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. *International Journal of Management Reviews* 19, 4 (2017), 432–454.
[4] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics.

[10] https://github.com/serenpa/Blog-Credibility-Corpus

			TF/IDF				Topic Modelling				BERT			
Label Set	**Balanced Class Size**	**Total Data**	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
Reasoning + Conclusion	1808	3616	0.609	0.588	0.598	0.622	0.624	0.683	**0.652**	0.651	0.593	0.654	0.622	0.619
Reasoning + Conclusion + Claim	9646	19292	0.655	0.726	**0.689**	0.67	0.591	0.64	0.614	0.596	0.597	0.696	0.642	0.611
Reasoning	1586	3172	0.601	0.619	0.61	0.639	0.547	0.63	0.585	0.594	0.607	0.685	**0.644**	0.655
Conclusion	331	662	0.625	0.583	**0.603**	0.654	0.471	0.533	0.5	0.519	0.469	0.633	0.539	0.511
Claim	9202	18404	0.634	0.698	**0.664**	0.644	0.578	0.632	0.604	0.582	0.593	0.655	0.622	0.6
Evidence (Citation/ code/events/experience/ref/stats/ Other)	3636	7272	0.665	0.597	**0.629**	0.661	0.551	0.636	0.591	0.576	0.58	0.587	0.583	0.597
Code Snippet	61	122	0.75	0.6	0.667	0.76	0.75	0.3	0.429	0.68	0.727	0.8	**0.762**	0.8
Events	261	522	0.718	0.571	0.636	0.695	0.531	0.531	0.531	0.562	0.656	0.816	**0.727**	0.714
Experience	2590	5180	0.711	0.638	**0.672**	0.69	0.612	0.63	0.621	0.617	0.617	0.605	0.611	0.616
Reference to Table/Image	29	58	1	0.333	0.5	0.667	0	0	0	0.25	0.571	0.667	**0.615**	0.583
Statistics/Data	22	44	1	0.333	0.5	0.556	0	0	0	0.333	1	0.833	**0.909**	0.889
Other	29	58	0.75	0.5	0.6	0.667	0	0	0	0.333	0.8	0.667	**0.727**	0.75
Citation	778	1556	0.597	0.537	0.565	0.606	0.481	0.517	0.498	0.503	0.522	0.557	**0.539**	0.545

Table 6: Preliminary results of detecting other labels within the dataset using Random Forests

In *Proceedings of the first workshop on argumentation mining*. Association for Computational Linguistics, Baltimore, Maryland USA, 64–68.

[5] Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, and Gurpreet Kaur. 2016. Opinion mining and sentiment analysis. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. ACM, New Delhi, India, 452–455.

[6] Marcus Banks. 2009. Blog posts and tweets: the next frontier for grey literature. In *Grey literature in library and information studies*. De Gruyter, UCSF Library, USA.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[8] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and regression trees*. CRC press, Boca Raton, Florida.

[9] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 1–27.

[10] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 670–680. https://www.aclweb.org/anthology/D17-1070

[11] 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, New York, NY, USA, 108–119.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[13] Nicola Dragoni, Saverio Giallorenzo, Alberto Lluch Lafuente, Manuel Mazzara, Fabrizio Montesi, Ruslan Mustafin, and Larisa Safina. 2017. *Microservices: Yesterday, Today, and Tomorrow*. Springer International Publishing, Cham, 195–216. https://doi.org/10.1007/978-3-319-67425-4_12

[14] Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. The COLING 2016 Organizing Committee, Osaka, Japan, 76–84. https://www.aclweb.org/anthology/W16-4011

[15] 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, New York, NY, USA, 1–6.

[16] 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.

[17] Jacob Goldberger, Geoffrey E Hinton, Sam T. Roweis, and Russ R Salakhutdinov. 2005. Neighbourhood Components Analysis. In *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou (Eds.). MIT Press, Cambridge, MA, 513–520. http://papers.nips.cc/paper/2566-neighbourhood-components-analysis.pdf

[18] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, 3483–3487.

[19] James Lewis and Martin Fowler. 2014. Microservices. https://martinfowler.com/articles/microservices.html

[20] 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.

[21] Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)* 16, 2 (2016), 1–25.

[22] Marco Lippi and Paolo Torroni. 2016. MARGOT: A web server for argumentation mining. *Expert Systems with Applications* 65 (2016), 292–303.

[23] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the Variance of the Adaptive Learning Rate and Beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*. OpenReview, Ethiopia, na.

[24] Ericka Menchen-Trevino and Eszter Hargittai. 2011. YOUNG ADULTS'CREDIBILITY ASSESSMENT OF WIKIPEDIA. *Information, Communication & Society* 14, 1 (2011), 24–51.

[25] Jana Novovičová, Antonín Malík, and Pavel Pudil. 2004. Feature selection using improved mutual information for text classification. In *Joint IAPR International*

Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, Springer, Berlin, Heidelberg, 1010–1017.

[26] 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, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[27] John C. Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, Cambridge, MA, 61–74.

[28] Austen Rainer. 2017. Using argumentation theory to analyse software practitioners' defeasible evidence, inference and belief. *Information and Software Technology* 87 (2017), 62–80.

[29] 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.

[30] 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.

[31] Austen Rainer, Ashley Williams, Vahid Garousi, and Michael Felderer. 2020. Retrieving and mining professional experience of software practice from grey literature: an exploratory review. *IET Software* 14, 6 (2020), 665–676.

[32] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[33] Zaher Salah, Abdel-Rahman F Al-Ghuwairi, Aladdin Baarah, Ahmad Aloqaily, Bar'a Qadoumi, Momen Alhayek, and Bushra Alhijawi. 2019. A systematic review on opinion mining and sentiment analysis in social media. *International Journal of Business Information Systems* 31, 4 (2019), 530–554.

[34] 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.

[35] Jacopo Soldani, Damian Andrew Tamburri, and Willem-Jan Van Den Heuvel. 2018. The Pains and Gains of Microservices: A Systematic Grey Literature Review. *Journal of Systems and Software* 146 (2018), 215–232.

[36] 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.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, NY, 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[38] Ashley Williams. 2018. Do software engineering practitioners cite research on software testing in their online articles? A preliminary survey.. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*. ACM, New York, NY, USA, 151–156.

[39] Ashley Williams. 2018. Using reasoning markers to select the more rigorous software practitioners' online content when searching for grey literature. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*. ACM, New York, NY, USA, 46–56.

[40] Ashley Williams and Austen Rainer. 2016. Identifying practitioners' arguments and evidence in blogs: insights from a pilot study. In *2016 23rd Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, Hamilton, NZ, 345–348.

[41] Ashley Williams and Austen Rainer. 2017. *The analysis and synthesis of previous work on credibility assessment in online media: technical report.* Technical Report. University of Canterbury, NZ.

[42] Ashley Williams and Austen Rainer. 2019. Do software engineering practitioners cite software testing research in their online articles? A larger scale replication. In *Proceedings of the Evaluation and Assessment on Software Engineering*. ACM, New York, NY, USA, 292–297.

[43] 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, New York, NY, USA, 211–220.

[44] Olaf Zimmermann. 2017. Microservices tenets. *Computer Science-Research and Development* 32, 3 (2017), 301–310.