Effect of Combination of HBM and Certainty Sampling on Workload of Semi-Automated Grey Literature Screening

Jinghui Lu [1] **Maeve Henchion** [2] **Brian Mac Namee** [1]

Abstract

With the rapid increase of unstructured text data, grey literature has become an important source of information to support research and innovation activities. In this paper, we propose a novel semi-automated grey literature screening approach that combines a Hierarchical BERT Model (HBM) with active learning to reduce the human workload in grey literature screening. Evaluations over three real-world grey literature datasets demonstrate that the proposed approach can save up to 64.88% of the human screening workload, while maintaining high screening accuracy. We also demonstrate how the use of the HBM model allows salient sentences within grey literature documents to be selected and highlighted to support workers in screening tasks.

1. Introduction

Grey literature is information that is either unpublished or published in non-commercial form (Farace & Frantzen, 2004). For example, government reports, newsletters, policy statements and unpublished research are considered grey literature, while advertisements and published peer-reviewed articles are not considered grey literature.

Recently, grey literature relating to research and innovation in the form of digitalised text is rapidly increasing in volume and accessibility, leading more and more innovation scholars to explore and exploit grey literature information to support research and innovation activities (Godin et al., 2015; Adams et al., 2016). However, a systematic grey literature review is costly, and screening grey literature to select documents relevant to the subject of the review is the most time-consuming part. This paper proposes a semi-automated method to reduce the human workload in grey literature screening. We frame semi-automated grey literature screening as a problem of using active learning for grey literature classification. More specifically, we develop a predictive model using active learning to assign relevant/irrelevant labels to grey literature documents. To the best of our knowledge, there is no prior work specifically related to automated or semi-automated grey literature screening.

There are, however, a number of relevant examples in the literature of using active learning to reduce the screening workload of academic papers especially in the medical domain. Most prior work in semi-automated academic citation screening aims to optimise selection strategies to minimise the number of documents reviewed, such as uncertainty sampling (Miwa et al., 2014; Saha et al., 2016; Ros et al., 2017; Varghese et al., 2019), certainty sampling (Miwa et al., 2014; Cormack & Grossman, 2016; Saha et al., 2016), patient active learning (Wallace et al., 2010b), or some more sophisticated selection strategies that mix the advantages of several methods (GROSSMAN & Cormack, 2016; Yu et al., 2018; Nghiem & Ananiadou, 2018). More recently, Yu et al. (2018) proposed the FASTREAD strategy that is a hybrid of several previous methods including weighting (Miwa et al., 2014) and aggressive undersampling (Wallace et al., 2010b), and has been demonstrated to outperform other previous baseline approaches on three software engineering datasets with respect to reducing screening workload.

A number of studies also investigate the impact of different text representations on systems developed for semi-automated literature screening. For example, combining SVM with doc2vec, word2vec, and LDA (Miwa et al., 2014; Hashimoto et al., 2016; Siddhant & Lipton, 2018). Hashimoto et al. (2016) propose a method, Paragraph Vector-based Topic Detection (PV-TD), that combines doc2vec (Le & Mikolov, 2014) (an extension of word2vec) with k-means clustering to perform simple topic modelling. PV-TD was shown to perform well compared to representations based on LDA and word2vec. Interestingly, Singh et al. (2018) extend the experiments in Hashimoto et al. (2016) with more datasets in the health domain, demonstrating that directly using doc2vec or bag-of-words (BOW) representations, rather

[1]School of Computer Science, University College Dublin, Dublin, Ireland [2]Teagasc Agriculture and Food Development Authority, Dublin, Ireland. Correspondence to: Jinghui Lu <Jinghui.Lu@ucdconnect.ie>, Brian Mac Namee <Brian.MacNamee@ucd.ie>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

than PV-TD, can achieve better results.

Though there are many similarities between semi-automated grey literature screening and academic literature screening, some differences exist, and hence more exploration is required in semi-automated grey literature screening. For example, in a systematic literature review for evidence-based practice (an approach to deliver health care by synthesising knowledge from studies and clinical data), practitioners cannot afford to miss any relevant documents because that means missing a potential treatment (Choi et al., 2012). However, in a systematic grey literature review for research and innovation, researchers aim to find new and useful information (Godin et al., 2015) rather than to evaluate all information related to a topic. So, it is acceptable to miss some relevant documents in a grey literature review. Also, the nature of grey literature datasets, such as the lack of document structure (grey literature lacks keywords, abstracts, etc.), necessitates specifically tailored approaches to be used.

With recent advances in text representations provided by BERT (Devlin et al., 2018) and its variants (Liu et al., 2019; Lan et al., 2019; Sanh et al., 2019), active learning systems have further improved the performance for text classification. Lu & MacNamee (2020) proposed the Adaptive Tuning Active Learning (ATAL) process where fine-tuning was incorporated into an active learning procedure, demonstrating that active learning could benefit from the iteratively improved document representations by fine-tuning language models with the labelled data in hand. We believe that it is possible to take advantage of BERT-like models to form novel document representations that are suitable for grey literature.

In this work, we argue that applying Hierarchical BERT Model (HBM) (Lu et al., 2021) — a BERT variant that considers the sentence-level information — can alleviate problems caused by the characteristics (i.e. lack of document structure) of data from grey literature. Then, we propose combining the HBM with active learning to build a semi-automated system to reduce screening workload. Certainty-based active learning, which has been shown to be very effective in semi-automated screening of academic literature (Miwa et al., 2014; Hashimoto et al., 2016; Przybyła et al., 2018). Hence, we propose applying HBM+Certainty-based active learning to semi-automate the screening of grey literature.

Also, Wallace et al. (2011) point out that annotation efficiency is an obstacle in active learning, and sentence highlighting as an extra benefit of HBM (noting that this was established in Lu et al. (2021)) can provide explanatory sentences which are helpful in the annotation tasks that must be completed by a human annotator. This work shows that HBM+active learning can select explanatory sentences to support a human reviewer's work.

Our contributions are: (1) we are the first to propose a solution to the task of semi-automated grey literature screening; (2) this paper presents several experiments based on three real-world agri-food grey literature datasets to demonstrate the utility of HBM+Certainty-based active learning in reducing the manual workload in grey literature screening.

The remainder of the paper is organised as follows: Section 2 describes the Hierarchical BERT Model (HBM); Section 3 describes experiments to demonstrate the effectiveness of HBM+Certainty in semi-automated screening of grey literature; Section 4 presents and discusses the experimental results and Section 5 demonstrates that the selected explanatory sentences can be a helpful support for human screeners; finally, Section 6 draws conclusions.

2. The Hierarchical BERT Model

Inspired by the Hierarchical Attention Network (HAN) (Yang et al., 2016), which was the first work to use sentence-level information for text classification, Lu et al. (2021) proposed the Hierarchical BERT Model (HBM) that extends BERT to use sentence structure information in forming document representations. The model consists of 3 components: (1) the token-level Roberta encoder (Liu et al., 2019) that extracts the token-level features and forms the vector representation for each sentence; (2) the sentence-level BERT encoder that takes the sentence vectors generated by the token-level Roberta encoder as inputs, generating the document representation by considering the association between sentences of a document; and (3) the prediction layer that predicts the class of a document based on the document representation provided by the sentence-level BERT encoder. The architecture and detailed implementations of each component are described in Lu et al. (2021).

Lu et al. (2021) conducted various evaluation experiments on six datasets of different sources (e.g. online reviews, blogs, news articles), demonstrating that HBM works well in low labelled data scenario (i.e. size of training set ranges from 50 to 200). Also, they proposed the method to infer important sentences based on the attention weights inside the HBM, and conducted a user study to show that the important sentences were helpful in accelerating human labelling tasks. Therefore, we believe that the combination of HBM and active learning can be effective in reducing the human workload in grey literature screening. In the next section, we will present experiments that demonstrate the utility of combining HBM with active learning for semi-automated grey literature screening.

3. Evaluating the HBM+Certainty Sampling for Semi-automated Grey Literature Screening

In this section we describe a set of experiments designed to evaluate the effectiveness of the HBM+Certainty approach to semi-automated grey literature screening. The following subsections describe the datasets used in the experiments, the active learning framework, the baseline approaches used in the experiments, and the evaluation metrics used.

3.1. Datasets

We evaluate the performance of various active learning methods on three fully labelled (i.e. relevant/irrelevant) agri-food domain grey literature datasets. The use of fully labelled datasets allows us to simulate data labelling by a human oracle, and is common in active learning for semi-automated academic citation screening (Hu et al., 2010; Hashimoto et al., 2016; Zhang et al., 2017; Zhao, 2017; Singh et al., 2018). These datasets are collected and labelled by agri-food domain experts working in the area of agri-food research and innovation from the Teagasc research centre.[1]

- **Animal By-Products:** 152 *relevant* articles and 962 *irrelevant* articles regarding applications of animal by-products. Articles about animal by-products applications are labelled as relevant, while articles that mention by-products but do not describe any specific application (i.e. animal by-products news, policies and so on) are labelled as irrelevant. The average number of sentences per document is 96.16 and the maximum number of sentences is 7974.
- **Anaerobic Digestion:** 223 *relevant* articles and 533 *irrelevant* articles regarding anaerobic digestion. The articles concerning current and future anaerobic digestion inputs and technologies are labelled as relevant articles, while the irrelevant ones are mainly relating to anaerobic digestion policy, community and economics, as well as advertisements. The average number of sentences per document is 58.74 and the maximum number of sentences is 2164.
- **Mastitis:** 162 *relevant* articles and 375 *irrelevant* articles regarding mastitis in cows. The articles concerning dairy cow mastitis are labelled as relevant articles, while irrelevant ones are mainly relating to other animal mastitis or other infections such as mad cow disease. The average number of sentences per document is 57.19 and the maximum number of sentences is 875.

It should be noted that the documents collected are in different formats such as PDFs, Word documents, and html files. For Word documents and html files, we can directly extract the raw text. For PDFs, optical character recognition tools are used to extract text information.[2]

3.2. Active Learning Framework

We apply pool-based active learning (Settles, 2009). At the outset, we train all classifiers with the same 10 documents (i.e. 5 positive documents and 5 negative documents) sampled at random from a dataset to seed the initial labelled data pool $\mathcal{L}$ for the active learning process. Subsequently, 10 unlabelled documents, whose ground truth labels will be revealed to each classifier, are selected according to a certain selection strategy. These documents are moved from the unlabelled data pool $\mathcal{U}$ to $\mathcal{L}$ (with their labels) and the classifiers are retrained based on the current $\mathcal{L}$. After each round the retrained classifier is used to label all of the examples remaining in $\mathcal{U}$ and we evaluate the performance of the system at this point. We repeat this active learning procedure until all documents are labelled. Also, the active learning process is repeated 10 times using different random seeds and the performance measures reported are averaged across these repetitions.

3.3. Baselines and Setup

We compare the HBM+Certainty approach with several baseline approaches. These baseline approaches are structured as a combination of a text representation technique commonly used in automated citation screening studies, a state-of-the-art selection strategy from other domains (i.e. screening social science/software engineer/biomedical literature), and a classifier. Unless otherwise stated the classifier used is an SVM as this has been shown to be effective in active learning for text classification problems (Wallace et al., 2010a; Miwa et al., 2014; Hashimoto et al., 2016; Singh et al., 2018).

Uncertainty sampling, which is generally regarded as a good selection strategy from many domains such as evidence-based medicine (Ma, 2007; Wallace et al., 2010b; Carvallo & Parra, 2019) and clinical text (Chen et al., 2012), is adopted as a baseline selection strategy in the experiment. Here, we combine uncertainty sampling with various text representations based on HBM[3], Roberta[4] and fine-tuned Roberta using the ATAL scheme described in (Lu & MacNamee, 2020)[5] (which are denoted as Uncertainty+HBM, Uncer-

[1] https://www.teagasc.ie/

[2] We used the BeautifulSoup (https://pypi.org/project/beautifulsoup4/) and pdfminer (https://pypi.org/project/pdfminer/) package for parsing and extracting text from htmls and pdfs.

[3] HBM method adopts a fine-tuned HBM model with a softmax layer as a classifier.

[4] We used the "roberta-base" model which can be found on https://github.com/huggingface/transformers.

[5] The ATAL approach adopts a fine-tuned Roberta model with

tainty+Roberta and Uncertainty+ATAL).

Certainty sampling, which is another good selection strategy for semi-automated academic literature screening (Miwa et al., 2014; Cormack & Grossman, 2016; Saha et al., 2016), is also adopted as a baseline selection strategy in the experiment. Also, we combine certainty sampling with various text representations based on HBM, Roberta and fine-tuned Roberta using the ATAL scheme (which are denoted as Certainty+HBM, Certainty+Roberta and Certainty+ATAL).

FASTREAD (Yu et al., 2018), a state-of-the-art selection strategy proposed for screening software engineering citations, is adopted as another baseline selection strategy. Yu et al. (2018) used TF-IDF in FASTREAD, however, based on the preliminary experiments we found that TF-IDF can not achieve results that are comparable with other deep-learning-based text representations. Hence, TF-IDF was also replaced with representations based on HBM, Roberta and ATAL (which are denoted as FASTREAD+HBM, FASTREAD+Roberta and FASTREAD+ATAL).

Hashimoto et al. (2016) proposed the Paragraph Vector Topic Detection+Certainty (Certainty+PV-TD) method which has been shown to outperform a word2vec-based method in health citations screening. This is adopted as another baseline approach in our experiment.

For baseline approaches using an SVM classifier (i.e. Roberta/ATAL/PV-TD-based approaches), we adopt the SVM+weighting scheme to counter the imbalance problem, using settings following Miwa et al. (2014). The parameters used for PV-TD follow those used by Hashimoto et al. (2016). The settings of ATAL are following Lu & MacNamee (2020), except that we start the first fine-tuning when 50 instances are labelled based on some preliminary experiments. For all HBM-based methods, all hyper-parameter settings are following Lu et al. (2021), except that we set the maximum number of sentences to 512, 512 and 128 for the *Animal by-products*, *Anaerobic Digestion*, and *Mastitis* datasets respectively based on the document lengths of these datasets and use 15 epochs for training.

3.4. Evaluation Metrics

As highlighted previously, what innovation scholars want from a grey literature review is to find a large number of relevant documents. This is in contrast to the goal of typical academic literature review screening in which finding all sources is key. In a typical systematic academic literature review, researchers cannot afford to miss any potential approaches, especially in the medical field. In this context, various evaluation metrics highlighting recall are usually used in semi-automated screening of academic literature, for example Yield, U19 (Wallace et al., 2010a), and F3 (Bekhuis & Demner-Fushman, 2012). However, since we do not emphasise recall in semi-automated grey literature screening, other metrics are more appropriate.

Based on the assumption that, in practice, users will have limited time during active learning which is quite common in a grey literature review (users will cease the procedure when they think they have reviewed enough data), we consider that a good system can provide more relevant documents as early as possible. Hence, we use coverage (Miwa et al., 2014)[6] to measure the performance of the different systems examined. Coverage measures the ratio of the relevant documents reviewed by the oracle (i.e. the human annotator) so far to the total relevant documents in the dataset. More specifically, coverage can be defined as follows:

$$\text{coverage} = \frac{TP^H}{TP^H + TP^M + FN^M} \tag{1}$$

where TP^H denotes the number of *true positives* reviewed by the oracle; and TP^M and FN^M indicate the number of *true positives* and *false negatives* predicted by the machine respectively. Intuitively, for a given manual annotation workload, the higher the coverage performance the better the active learning method.

Also, by computing the Area Under the Coverage Curve (AUCC), we can estimate how quickly the system can provide the user with relevant documents (Przybyła et al., 2018). To further understand the AUCC, we can view Figure 1 which shows the coverage performance of the Certainty+HBM method when applied to the Animal Byproducts dataset. In this figure, the X-axis is the number of labelled documents and the Y-axis represents performance measured using coverage. The area coloured by light blue denotes the AUCC score of the Certainty+HBM curve. Intuitively, a larger area indicates better performance.

For evaluating the workload saved by the active learning system, we use the work saved over sampling at 95% coverage (referred to as WSS@95%) (Przybyła et al., 2018), which indicates the percentage of documents that the reviewer does not have to read (because those documents has been safely screened out by the active learning system) when the active learning system yields a coverage performance of 95% (i.e. 95% of relevant documents has been reviewed by the oracle)[7], compared with screening in a random order. WSS@95% can be calculated by the following equation:

a softmax layer as a classifier.

[6] Przybyła et al. (2018) also rephrase this metric as recall in the active learning context, to avoid confusion, we use the name coverage in this paper.

[7] In practice, it is not possible for a reviewer to know exactly when the desired coverage has been achieved. Here we use the fully-labelled dataset to simulate the data labelling process.

$$\text{WSS@95\%} = \frac{0.95N - X_{95}}{N} \quad (2)$$

where X_{95} is the number of documents that have been reviewed by the oracle when the active learning system achieves coverage of 95%; and N is the total number of documents in the dataset. We assume that if random selection were used, an oracle would have to label 95% of the documents in a pool to achieve 95% coverage. Thus, the reduction between $0.95N$ and X_{95} is the workload saved by the active learning system to be evaluated. The saved workload is normalised by dividing the total number of documents, N. Intuitively, a larger WSS@95% indicates a better active learning system.

Figure 1 also illustrates the WSS@95% metric. The horizontal coordinate of the dashed black vertical line is 345, which indicates that the Certainty+HBM model reaches coverage of 95% after 345 documents have been labelled by the oracle. Hence, the WSS@95% of Certainty+HBM is approximated by (0.95*1114 - 345)/1114 $\approx$ 64.03% (the number of total documents in the *Animal By-products* dataset is 1114).

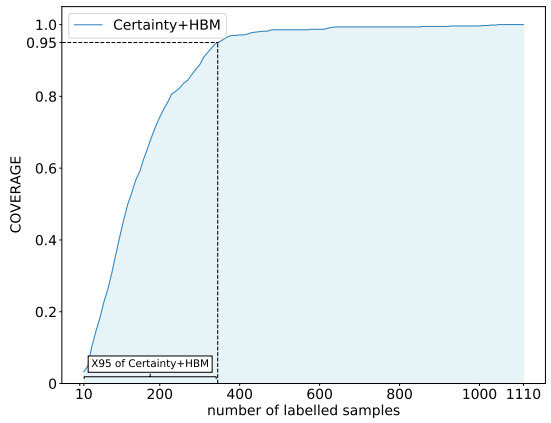

Figure 1. The coverage performance of the Certainty+HBM method when applied to the *Animal By-products* dataset. The X-axis represents the number of documents that have been manually annotated and the Y-axis denotes the coverage. The curve starts with 10 along the X-axis because of the seeded instances. The horizontal coordinate of the dashed vertical line is X_{95} of the coverage curve, which indicates the number of documents that have been manually annotated when an optimal coverage performance of 95% is reached. The area coloured by light blue denotes the AUCC score of the coverage curve.

4. Results and Discussion

This section presents and discusses the results of the experiments run to evaluate the effectiveness of the HBM approach to semi-autoamted grey literature screening.

4.1. Results

Figures 2, 3 and 4 show the performance, measured using coverage, achieved by various active learning systems when applied to the *Animal By-products*, *Anaerobic Digestion* and *Mastitis* datasets respectively. In these figures, the X-axis is the number of labelled documents and the Y-axis is the performance measured using coverage. In all cases, the coverage curves dramatically increase, which can be explained by the fact that as the number of reviewed documents increases there are more and more relevant documents included and reviewed by the oracle.

For example, in the *Animal By-products* dataset (see Figure 2), if given a specific manual annotation workload of 20% (i.e. 20% of the documents are manually reviewed), HBM+Certainty seems to be the best method which discovers 79.04% of relevant documents compared to 75.71% of relevant documents retrieved by ATAL+Certainty (the second best performing method), and 68.57% by Roberta+Certainty (the third best performing method).

With regard to the *Anaerobic Digestion* dataset (see Figure 3), the coverage performance of HBM+Certainty and ATAL+Certainty are very close and are better than that of other methods. For the *Mastitis* dataset (see Figure 4), during all iterations, we observe a similar pattern that the performance obtained by HBM+Certainty is slightly higher than that of ATAL+Certainty and these two methods outperform other methods by a reasonable margin.

We can also look at Table 1 which summarises the performance of the active learning methods compared with respect to AUCC score to further assess the effectiveness of each method. In this table, the first column denotes active learning methods. Each row denotes the AUCC score of the specific method over three different grey literature datasets. The AUCC scores are calculated by the trapezoidal rule and normalised by the maximum possible area, to bound the values between 0 and 1. The best performing AUCC score of each dataset is highlighted and the number in parentheses indicates the rank of the method regarding a specific dataset (the smaller the number the higher the rank).

Overall, again, the experimental results demonstrate that HBM+Certainty is the most effective method in two grey literature datasets. However, in the *Anaerobic Digestion* dataset, ATAL+Certainty achieves a slight improvement in performance over the HBM+Certainty approach. By contrast, it seems that Certainty+PV-TD is the least effective method for screening grey literature as it performs worst in

Table 1. AUCC score of 10 different active learning methods over three grey literature datasets. The best performing AUCC score of each dataset is highlighted and the number in parentheses indicates the rank of the method regarding a specific dataset (the smaller the number the higher the rank).

Methods	Animal By-Products	Anaerobic Digestion	Mastitis	Avg Rank
FASTREAD+Roberta	0.8077(5)	0.8138(5)	0.5786(8)	6.0
FASTREAD+ATAL	0.8226(4)	0.8381(4)	0.6451(4)	4.0
FASTREAD+HBM	0.7841(8)	0.8109(6)	0.7040(3)	5.6
Certainty+Roberta	0.8353(3)	0.8394(3)	0.5748(9)	5.0
Certainty+ATAL	0.8657(2)	**0.8551(1)**	0.7133(2)	1.6
Uncertainty+Roberta	0.8056(6)	0.4851(10)	0.6239(5)	7.0
Uncertainty+ATAL	0.7022(9)	0.7519(7)	0.6022(7)	7.6
Uncertainty+HBM	0.7991(7)	0.6150(8)	0.6119(6)	7.0
Certainty+PV-TD	0.6459(10)	0.5387(9)	0.4957(10)	9.6
Certainty+HBM	**0.8713(1)**	0.8491(2)	**0.7394(1)**	1.3

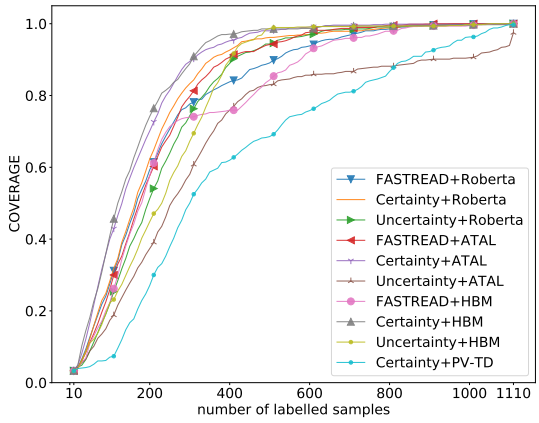

Figure 2. Comparisons of performance of 10 different active learning processes when applied to the *Animal By-products* dataset.

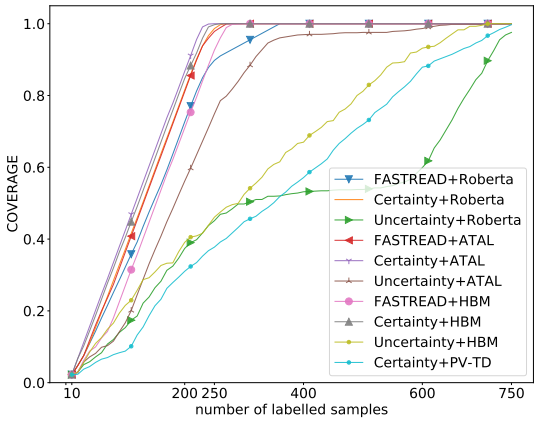

Figure 3. Comparisons of performance of 10 different active learning processes when applied to the *Anaerobic Digestion* dataset.

two out of three grey literature datasets.

Figure 5 summarises the WSS@95% scores obtained by the 10 active learning systems compared across three agri-food domain grey literature datasets, which can be used to assess the different approaches in terms of work saved. It can be noted that, with respect to WSS@95%, the HBM+Certainty approach outperforms other methods by a large margin in the *Animal By-products* dataset. The improvements varied from 4.34% compared with ATAL+Certainty to 68.03% compared with ATAL+Uncertainty. In the other two datasets, though the HBM+Certainty approach does not save the most workload, it still reaches a performance level close to the best performing methods, i.e. ATAL+Certainty for the *Anaerobic Digestion* dataset and HBM+FASTREAD for the *Mastitis* dataset. It is also interesting to discover that some methods achieve negative WSS@95% scores, which indicates the methods fail to reduce the manual annotation workload, e.g. Certainty+PV-TD in the *Mastitis* dataset.

4.2. Discussion

These experimental results show that with respect to Area Under Coverage Curve (AUCC), the proposed HBM+Certainty approach surpasses the state-of-the-art methods from other domains (i.e. PV-TD and FASTREAD), demonstrating the effectiveness of the proposed method when applied to semi-automated screening of grey literature.

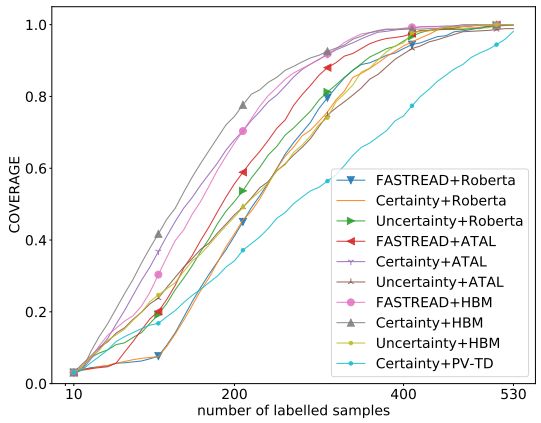

Figure 4. Comparisons of performance of 10 different active learning processes when applied to the *Mastitis* dataset.

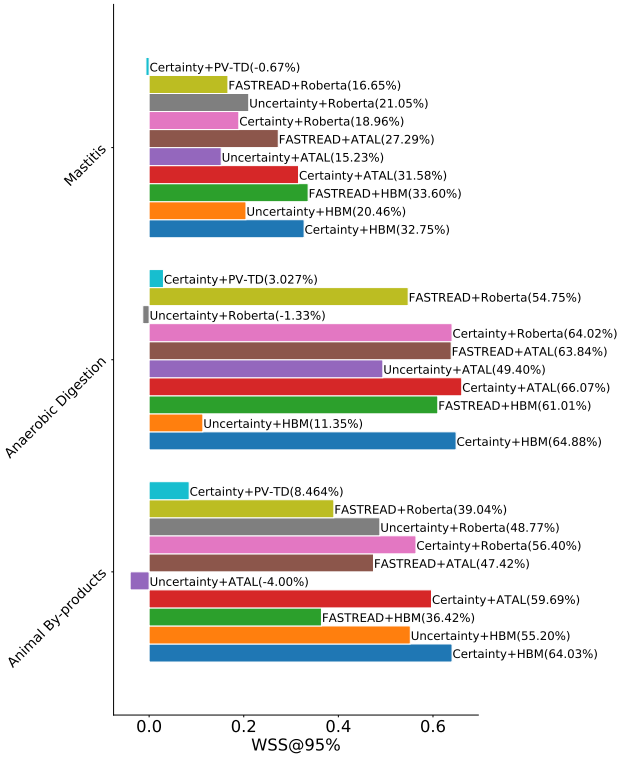

Figure 5. WSS@95% achieved by different active learning methods across three agri-food domain grey literature datasets. The horizontal axis denotes the WSS@95%. The method name and its performance (i.e. number in brackets) are beside the corresponding horizontal bar.

Even though with respect to WSS@95%, HBM+Certainty cannot consistently outperform other methods, it is more likely due to the different focuses of the two evaluation metrics. WSS@95% focuses on the workload at the fixed moment when 95% of relevant documents are retrieved, the AUCC metric averages the workload across all coverage levels. It is likely that usually the reviewer will end the active learning process much earlier than when 95% of relevant documents are retrieved when screening grey literature in practice. Hence, we focus more on the AUCC metric which reflects the overall performance of the active learning system.

Also, some methods achieve negative WSS@95% scores, which indicates that the methods fail to reduce the manual annotation workload (e.g. Certainty+PV-TD in the Mastitis dataset). This suggests that the active learning approach to semi-automated grey literature screening may not be a solution if not done correctly, hence highlighting the importance of the experiments undertaken here.

Consequently, we conclude that the proposed HBM+Certainty approach is an effective approach to semi-automated grey literature screening.

5. Highlighting Salient Sentences

As described in Lu et al. (2021), HBM, via the sentence-level self-attention mechanism, can be used to identify the important sentences in documents. These can be used as explanatory sentences to improve the efficiency of the oracle's labelling. In this section we provide an illustrative example to show how this would work in the active learning context with the agri-food datasets. The process for inferring important sentences is following Lu et al. (2021).

Figures 6 shows an example of a relevant article from the *Animal By-products* datasets.[8] The selected sentences are highly related to applications of animal by-products which is evidenced by the text *"A myriad of uses for these items – leather products from hides lubricants plastics soaps glycerin gelatins and ....."* and *"In the U.S. edible offal animal organs such as liver heart and stomach is used to produce sausages hot dogs and other processed meat products it is also a major ingredient in pet foods. "*. Also, the model successfully skipped irrelevant sentences such as URL links, and named entities (i.e. titles, organizations). These highlighted sentences can support a human screener to perform screening faster during the active learning process.

[8]More examples and code are available at https://github.com/GeorgeLuImmortal/Effect-of-Combination-of-HBM-and-Certainty-Sampling-on-Workload-of-Semi-Automated-Grey-Literature-Sc.

Beef and Pork Byproducts Enhancing the U.S. Meat Industry Bottom Line. Beef and pork production yields more than just what is seen on people plates. Byproducts—edible offal inedible offal blood hides and rendered products—in clude virtually all parts of the live animal that are not part of the dressed carcass. These items constitute an estimated percent of the liveweight of a hog and about per cent of the liveweight of cattle. A myriad of uses for these items—leather products from hides lubricants plastics soaps glycerin gelatins and other industrial household cosmetic pharmaceutical and medical supplies—allow the meat industry to capture additional revenue and avoid costs for dis posing of certain edible and nonedible parts of the animal. Exports and other markets for animal byproducts contribute to the value and prof itability of the meat processing industry and. ERS research indicates that a increase in the value of byproducts to processors adds about cents to the average price paid per hundredweight to producers of fed steers slaughter cattle that have been finished on concentrated feed on a per hundredweight basis. Conversely consumer prices for other beef products are lower than they would be without byproduct sales because the processing costs to whole salers of the entire animal are spread across both byproducts and meat. In the U.S. edible offal animal organs such as liver heart and stomach is used to produce sausages hot dogs and other processed meat products it is also a major ingredient in pet foods. In foreign markets demand for U.S. edible offal including variety meats edible byproducts that are segregated chilled and processed under. In U.S. exports of beef/veal and pork edible offal reached a record. Beef/veal and pork edible offal. Beef/veal and pork muscle meat cuts. Source USDA Economic Research Service using USDA Foreign Agricultural Service Global Agricultural Trade System data. sanitary conditions and are inspected for sanitation and wholesomeness by the U.S. Meat Inspection Service is high because of its superior quality and low prices rela tive to domestic products. Over the past years byproducts accounted for more than percent volume of U.S. beef and veal exports and percent volume of U.S. pork exports. Together edible beef/veal and pork byproduct exports account for more than percent of the value of total U.S. beef/veal and pork exports. In beef/veal and pork edible offal exports reached a record level of. billion million more than the previous record set in. Daniel L. Marti dmarti ers.usda.gov Rachel J. Johnson rjohnson ers.usda.gov Kenneth H. Mathews Jr. kmathews ers.usda.gov. This finding is drawn from. Variety Meat Exports and the Global Marketplace by Daniel L. Marti and Rachel J. Johnson in Livestock Dairy and Poultry Outlook LDP USDA Economic Research Service September available at www.ers.usda.gov/ publications/ldp//Sep/ldpm.pdf You may also be interested in. ERS Briefing Room on Cattle available at www.ers.usda.gov/briefing/cattle/ ERS Briefing Room on Hogs available at www.ers.usda.gov/briefing/hogs/

Figure 6. A relevant document concerning animal by-products applications in the *Animal By-products* dataset.

6. Conclusions

In order to achieve good performance in semi-automated screening of grey literature, a customised Hierarchical BERT Model combined with certainty-based active learning was proposed. We evaluated our approach against state-of-the-art active learning strategies from many domains based on three real-world agri-food grey literature datasets. Experimental results show that the proposed method achieved a great improvement with respect to coverage when compared to the baseline methods. Also, we demonstrated that the proposed method dramatically reduced the manual annotation cost while retaining 95% of relevant studies that were reviewed by the users. Also, by demonstrating the highlighted sentences identified by the model, we have shown the effectiveness of HBM+Certainty in selecting important sentences. In the future, we will design experiments to quantify the utility of the important sentences in helping human annotators with labelling grey literature. We will also explore the usage of HBM+Certainty sampling in the grey literature dataset of different domains such as medicine and politics.

Acknowledgements

This research was kindly supported by a Teagasc Walshe Scholarship award (2016053) and Science Foundation Ireland (12/RC/2289_P2).

References

Adams, J., Hillier-Brown, F. C., Moore, H. J., Lake, A. A., Araujo-Soares, V., White, M., and Summerbell, C. Searching and synthesising 'grey literature'and 'grey

information'in public health: critical reflections on three case studies. *Systematic reviews*, 5(1):1–11, 2016.

Bekhuis, T. and Demner-Fushman, D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artificial intelligence in medicine*, 55(3):197–207, 2012.

Carvallo, A. and Parra, D. Comparing word embeddings for document screening based on active learning. In *BIRNDL@ SIGIR*, pp. 100–107, 2019.

Chen, Y., Mani, S., and Xu, H. Applying active learning to assertion classification of concepts in clinical text. *Journal of biomedical informatics*, 45(2):265–272, 2012.

Choi, S., Ryu, B., Yoo, S., and Choi, J. Combining relevancy and methodological quality into a single ranking for evidence-based medicine. *Information Sciences*, 214: 76–90, 2012.

Cormack, G. V. and Grossman, M. R. Engineering quality and reliability in technology-assisted review. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 75–84, 2016.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Farace, D. and Frantzen, J. Sixth international conference on grey literature: work on grey in progress. In *Grey literature 2004 conference proceedings Amsterdam: TextRelease*, 2004.

Godin, K., Stapleton, J., Kirkpatrick, S. I., Hanning, R. M., and Leatherdale, S. T. Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in canada. *Systematic reviews*, 4(1):138, 2015.

GROSSMAN, M. R. and Cormack, G. Continuous active learning for tar. *The Journal*, 2016.

Hashimoto, K., Kontonatsios, G., Miwa, M., and Ananiadou, S. Topic detection using paragraph vectors to support active learning in systematic reviews. *Journal of biomedical informatics*, 62:59–65, 2016.

Hu, R., Delany, S. J., and Mac Namee, B. Egal: Exploration guided active learning for tcbr. In *International Conference on Case-Based Reasoning*, pp. 156–170. Springer, 2010.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196, 2014.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lu, J. and MacNamee, B. Investigating the effectiveness of representations based on pretrained transformer-based language models in active learning for labelling text datasets. *arXiv preprint arXiv:2004.13138*, 2020.

Lu, J., Henchion, M., Bacher, I., and MacNamee, B. A sentence-level hierarchical bert model for document classification with limited labelled data. *arXiv preprint arXiv:2106.06738*, 2021.

Ma, Y. *Text classification on imbalanced data: Application to Systematic Reviews Automation*. PhD thesis, University of Ottawa (Canada), 2007.

Miwa, M., Thomas, J., O'Mara-Eves, A., and Ananiadou, S. Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, 51: 242–253, 2014.

Nghiem, M.-Q. and Ananiadou, S. Aplenty: annotation tool for creating high-quality datasets using active and proactive learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 108–113, 2018.

Przybyła, P., Brockmeier, A. J., Kontonatsios, G., Le Pogam, M.-A., McNaught, J., von Elm, E., Nolan, K., and Ananiadou, S. Prioritising references for systematic reviews with robotanalyst: a user study. *Research synthesis methods*, 9(3):470–488, 2018.

Ros, R., Bjarnason, E., and Runeson, P. A machine learning approach for semi-automated search and selection in literature studies. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, pp. 118–127, 2017.

Saha, T. K., Ouzzani, M., Hammady, H. M., Elmagarmid, A. K., Dhifli, W., and Hasan, M. A. A large scale study of svm based methods for abstract screening in systematic reviews. *arXiv preprint arXiv:1610.00192*, 2016.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Settles, B. Active learning literature survey. 2009.

Siddhant, A. and Lipton, Z. C. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. *arXiv preprint arXiv:1808.05697*, 2018.

Singh, G., Thomas, J., and Shawe-Taylor, J. Improving active learning in systematic reviews. *arXiv preprint arXiv:1801.09496*, 2018.

Varghese, A., Hong, T., Hunter, C., Agyeman-Badu, G., and Cawley, M. Active learning in automated text classification: a case study exploring bias in predicted model performance metrics. *Environment Systems and Decisions*, 39(3):269–280, 2019.

Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 173–182. ACM, 2010a.

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., and Schmid, C. H. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11 (1):1–11, 2010b.

Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. Who should label what? instance allocation in multiple expert active learning. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pp. 176–187. SIAM, 2011.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480–1489, 2016.

Yu, Z., Kraft, N. A., and Menzies, T. Finding better active learners for faster literature reviews. *Empirical Software Engineering*, 23(6):3161–3186, 2018.

Zhang, Y., Lease, M., and Wallace, B. C. Active discriminative text representation learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Zhao, W. Deep active learning for short-text classification, 2017.