Annotating Documents using Active Learning Methods for a Maintenance Analysis Application

James Pope
Mark Terwilliger
jpope6@una.edu
mterwilliger@una.edu
University of North Alabama
Alabama, United States

J.A. (Jim) Connell
Gabriel Talley
Nicholas Blozik
connellja@montevallo.edu
gtalley@forum.montevallo.edu
nblozik@forum.montevallo.edu
Stephens College of Business
University of Montevallo
Alabama, United States

David Taylor
dtylor11@memphis.edu
University of Memphis
Tennessee, United States

ABSTRACT

The aircraft cargo industry still maintains vast amounts of the maintenance history of aircraft components in electronic (i.e. scanned) but unsearchable images. For a given supplier, there can be hundreds of thousands of image documents only some of which contain useful information. Using supervised machine learning techniques has been shown to be effective in recognising these documents for further information extraction. A well known deficiency of supervised learning approaches is that annotating sufficient documents to create an effective model requires valuable human effort. This paper first shows how to obtain a representative sample from a supplier's corpus. Given this sample of unlabelled documents an active learning approach is used to select which documents to annotate first using a normalised certainty measure derived from a soft classifier's prediction distribution. Finally the accuracy of various selection approaches using this certainty measure are compared along each iteration of the active learning cycle. The experiments show that a greedy selection method using the uncertainty measure can significantly reduce the number of annotations required for a certain accuracy. The results provide valuable information for users and more generally illustrate an effective deployment of a machine learning application.

CCS CONCEPTS

• **Information systems** → *Data mining*; **Clustering and classification**.

KEYWORDS

Active Learning, Document Classification

AIPR 2020, June 26–28, 2020, Xiamen, China

ACM ISBN 978-1-4503-7551-1/20/06…$15.00
DOI: https://doi.org/10.1145/3430199.3430214

1 INTRODUCTION

A large, international airline company has currently amassed several million maintenance documents regarding aircraft components sent to third party vendors (termed suppliers). The supplier will diagnose and correct the fault, returning the component and associated maintenance documents to the airline. The maintenance document exchange was done via paper until the late 1990's when efforts began to transition to electronic exchange. However, two issues remain, the first being that not all suppliers have transitioned to electronic exchange. Second, most of the components whose maintenance history is in these inaccessible documents are on aircraft still in service. Thus, tracking the maintenance history of an aircraft typically requires a combination of electronic queries and manual searches through poorly scanned in documents. The manual searching is laborious and an automated search through the scanned in documents is required.

Analysing documents has been researched for some time and many tools exist for optical character recognition. To assist in automating the process, the airline needs a way to determine the document type. However, many problems hinder straight forward approaches including poor scanners, skew, misalignment and numerous document variants. Traditional supervised machine learning has been shown to address this issue [7]. However, during deployment two critical issues needed to be addressed.

- How many and which documents to select from a supplier's corpus?
- Given this sample and a selection approach, how many annotations are sufficient?

A known disadvantage of supervised learning is that it requires significant amounts of annotated observations. For our application, except for some meta-data (e.g. date scanned), the documents are unlabelled and need tedious human effort to annotate them. Typically active learning is used to select more informative observations from the population of unlabelled observations. In this paper, we use an active learning approach in a different way to assist in annotating a sample of unlabelled documents already selected from the population. This approach is depicted in Figure 1.

We use an uncertainty measure to select a batch of documents to show to an annotator for labelling. We compare the accuracy of various selection methods based on this measure for each active learning cycle. We demonstrate that a greedy approach of always

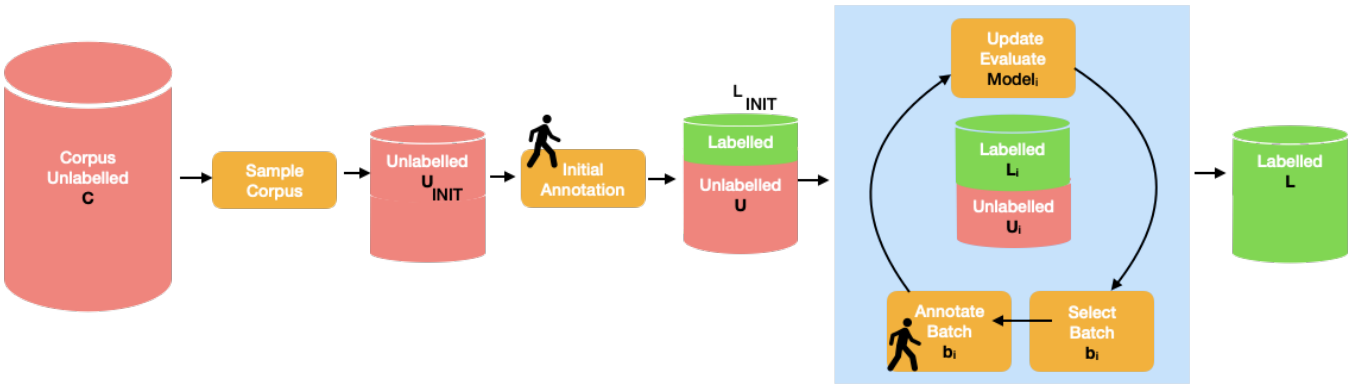

Figure 1: System Model

selecting a batch of the least certain documents more quickly leads to an effective model and can reduce the annotation effort by nearly four times. The contributions of the paper are as follows:

- Industry application of active learning to reduce human labour.
- Comparison of heuristic selection methods.
- Application of uncertainty measures and weighted random sampling for active learning.

The rest of the paper is organised as follows. The related work is discussed followed by the system model. The active learning cycle is then presented with the uncertainty measures and selection methods detailed. Experiments and results are then discussed followed by the conclusion.

2 RELATED WORK

Active learning has been researched seeking to address the labelled observation assumption in supervised machine learning. It can be seen as a special case of semi-supervised machine learning that mixes supervised and unsupervised learning with labelled and unlabelled data [12].

Nigam, et al. [5], propose leveraging unlabelled and labelled data for text classification using an expectation-maximisation approach. The approach also includes weighting factors and mixture component per class. Wang and Hua [11] provide a survey of active learning techniques. Their survey categorises approaches including those based on uncertainty measures.

Comparing models with multiple classes (multinomial) continues to be researched. The F1 measure is common to access the accuracy of individual classes as the harmonic mean of precision and recall. For multiple classes a simple average (usually termed *Macro F1*) or a weighted averaged of the F1 measures have been proposed [9, 13]. We primarily use the weighted F1 measure for comparisons.

Efraimidis [2] presents a reservoir sampling technique for selecting k items over a data stream using a weighting factor (i.e. weighted random sampling (WRS)). Some of our selection methods use this technique.

Lughofer [4] proposes a hybrid system in two phases using a combination of unsupervised learning in the first phase and an active learning approach using a certainty criterion. Our work was motivated separately but is most similar to this research.

3 SYSTEM MODEL

A text analysis pipeline has been established with the intent of identifying specific aircraft repair documents and extracting their text via optical character recognition for further processing. Figure 2 depicts the pipeline with the ultimate goal to determine repair specific actions (e.g. fault confirmed or not confirmed). The approach necessarily requires document type and orientation classification.

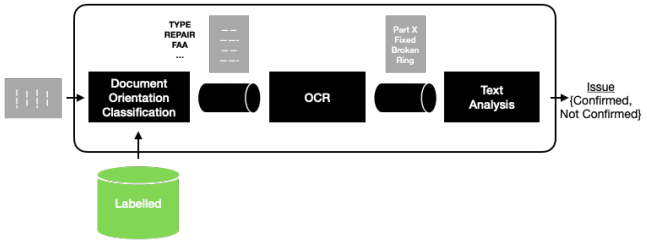

Figure 2: Application Pipeline

We are given a large corpus of documents from one supplier that have already been labelled with the document type and orientation. These were annotated by researchers and airline employee subject matter experts and constitute ground truth for our research. As depicted in Figure 1, given a new supplier's documents, users are expected to complete the following tasks:

- Take a sample of documents from the supplier's corpus.
- Annotate the document type and orientation of the sampled documents.
- Evaluate document type and orientation classifiers.
- Integrate the labelled documents into the pipeline.

3.1 Supervised Classification

In a previous work [7] we showed that roughly 145 features could be extracted from the documents and used to classify the document type and orientation with sufficient accuracy. The research showed that Support Vector Machines (SVM) [6, 8] and the ensemble Random Forest (RF) [1] classifiers provided superior accuracy for our approach. We restrict our assumption to only classifiers that can produce probability estimates for each class in their predictions

Table 1: Proportion by Class

Class	Count	Proportion
invoice	7522	0.2803
faa	5077	0.1892
tag	5076	0.1891
components	4540	0.1692
blank	2427	0.0904
other	1284	0.0478
receiving	419	0.0156
data	245	0.0091
repair	194	0.0072
exception	52	0.0019

(a.k.a. soft classifier). Our approach does not preclude other soft classifiers (e.g. Naive Bayes), however, for the remainder of the paper we assume the classifier to be the RF. We refer to the classifier with a specific set of annotated documents as a *model*. As annotated documents are added we assume a new model is produced.

3.2 Classifier Accuracy Measure

There are many ways to measure the accuracy of a classifier using a single number. For binary classifiers the harmonic mean between precision and recall is commonly used called the *F1* or *Fmeasure*. For multinomial classifiers the precision and recall for each class can first be determined. Next the average of the $F1$'s can be taken, usually termed the *Macro-F1* [9, 13], or a *Weighted F1* can be determined based on the class sizes. We believe, but do not show, that due to the class imbalance the *Weighted F1* is a better measure and compute it as follows:

$$\textit{Weighted F1} = \sum_{i=1}^{m} \frac{size(i)}{m} \times \frac{2 \times precision(i) \times recall(i)}{precision(i) + recall(i)} \quad (1)$$

where i represents the i'th class. The $size(i)$ function simply returns the number of instances in the class and m is the number of classes.

3.3 Corpus Sampling

There are 26,836 documents in the corpus with 10 classes (i.e. document types). Table 1 shows the distribution of these classes. Clearly there are large class imbalances where the first four are reasonably well represented in the corpus and the last four are more sparse. For the remainder of this paper we refer to this set of documents as the corpus.

There is metadata for the documents though mostly timeline information. The documents are roughly evenly spaced between the years 1996 and 2019. However, we do not believe there to be useful information in the metadata for a more direct active learning approach. Instead, we take a representative sample from the corpus from which to annotate and build a classifier (later we use active learning to select which documents in the sample to annotate first). More annotated samples will generally build a better classifier. We compare building classifiers using stratified sampling by year and from a simple random sample. From the sample we build an RF classifier and evaluate the accuracy against the corpus (acting as a test set) using the weighted F1 score. For the classifier implementation we use the data mining software described by Hall, et al. [3]. The experiments are repeated ten times with the average and 95% confidence interval results shown in Figure 3.

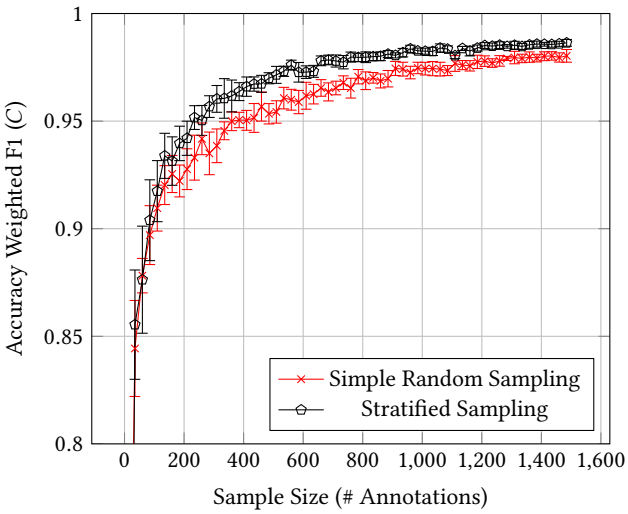

Figure 3: Sample Size vs Accuracy

The stratified sampling approach is slightly higher though in many cases within the margin of error. We conclude that there is no appreciable difference between the two approaches and that a simple random sample is sufficient. The figure also provides an idea of the accuracy possible for the corpus given some sample size with perhaps a 2-4% irreducible error. This answers the question concerning how to sample from the corpus. There are certainly differences between suppliers, however, the documents are still of the same nature (i.e. form style, with lines, text, symbols, etc.) and we submit that samples of this size will also generalise to the other supplier's corpus.

4 ACTIVE LEARNING CYCLE

Based on Section 3.3 we assume a simple random sample of 550 documents from the corpus. We denote this unlabelled set of documents as U. As suggested in Figure 1, the user needs to initially annotate U producing a small set of labelled documents denoted as L_{INIT}. For concreteness we found an initial size of 50 to be sufficient. We refer to the shaded region in Figure 1 as the active learning cycle where the user selects a set of documents (denoted the *batch*) from U, annotates them, moves to them to L, builds the model, and evaluates against the remaining U documents (treating similar to a validation set). The cycle is repeated until all documents have been annotated or the user sees evidence from the evaluation that the model is sufficient. Ideally users will not have to annotate the entire sample. We note that each iteration constitutes a new model.

There are numerous ways to implement the cycle. Unless specified otherwise, we assume a fixed $k = |batch|$ size of 10 for each iteration. For convenience, we will assume that all the documents

in the *batch* are to be labelled and that the user correctly annotates them. The user is presented with the *batch* of documents with suggested labels provided by the most recent model. We could estimate how well the *batch* improves the model and discard if it does not (similar to an expectation-maximisation (EM) approach). However, to simplify our exposition, we always take the selected *batch* and add to L. This means that it is possible for subsequent models to perform worse after adding a new *batch*.

Users would prefer a method that selects the most informative documents to be annotated before less informative documents. This intuitively results in more accurate models with fewer annotations. We choose to use an uncertainty measure restricting the classifier to those that can produce an estimated probability for each class. We do not detail how the user would know when to stop the cycle prior to annotating all of the samples but do show evidence that the uncertainty metric can be used to assist in making this determination. We first detail how the uncertainty measure is computed and then discuss how this measure is used to select from U.

4.1 Uncertainty Measure

We consider the distribution produced by the classifier for each of the m different document types. We treat the distribution as the uncertainty of the prediction. We take two or more document types with similar probabilities as evidence the classifier is uncertain about the prediction. In this paper we view entropy, denoted $H(X)$, as a measure of uncertainty. We define the uncertainty of the classifier's prediction as follows (taking notation from [10]).

$$H(X)_{pred} = \sum_{i=1}^{m} p(x_i) \log_2 \frac{1}{p(x_i)} \tag{2}$$

where we define X as a discrete random variable with, for our case, m being the number of different document types. $H(X)$ will be some positive number greater than or equal to zero. Note that $p(x_i) \log_2 \frac{1}{p(x_i)}$ is defined to be zero when x_i is zero. $H(X)$ is zero when one of the x_i's is 1.0 (i.e. the classifier is certain the target label is x_i). $H(X)$ can be greater than one and typically will be for large m. Uncertainty will be greatest when all $p(x_i)$ are uniformly distributed with $p(x_i) = \frac{1}{m}$. We define $H(X)_{max}$ as follows.

$$H(X)_{max} = \sum_{i=1}^{m} \frac{1}{m} \log_2 \frac{1}{\frac{1}{m}} = \log_2 m \tag{3}$$

We desire a bounded measure as trying to find some threshold entropy will vary between suppliers with differing number of document types. To address this, we normalise the predicted entropy by dividing by the maximum entropy.

$$H(X)_{norm} = \frac{H(X)_{pred}}{H(X)_{max}} \tag{4}$$

The normalised entropy $H(X)_{norm}$ will thus be in the range of 0.0 to 1.0. We interpret 1.0 as being the most uncertain the classifier can be about the prediction and 0.0 being the most certain. Let u_i be the computed normalised entropy when given the i'th document's prediction distribution from the classifier. For the remainder of the paper we will generally refer a document's uncertainty (or certainty) as follows:

$$\text{Document } i \text{ Uncertainty} \quad = u_i = H(x_i)_{norm}$$
$$\text{Document } i \text{ Certainty} \quad = c_i = 1 - H(x_i)_{norm}$$

4.2 Batch Selection Methods

We could present one document to the user at a time for annotation and update the model each time (i.e. incremental learning approach). To minimise model building without restricting to incremental classifiers, we choose to present a batch of $k = 10$ documents from the unlabelled set U for annotation to the user. We propose and detail the following selection methods.

4.2.1 LeastCertain. From the unlabelled set we sort ascending by c_i and then choose the first k documents that will have the least certainty. This would seem to prefer documents that are more difficult to classify to present to the user to annotate and may be more informative - conversely this method may learn faster.

4.2.2 MostCertain. From the unlabelled set we sort ascending by c_i and then choose the last k documents that will have the most certainty. Adding documents that the classifier is most certain about would not seem to add much information to the model and may take more iterations to learn. Towards the end of the iterations the more informative documents are likely still in the unlabelled set U. However, it is not clear that it would necessarily reduce the model's accuracy performance.

4.2.3 LessCertainWRS. We use a weighted random sample based on the uncertainty measure of the documents. The weighted random score for each document, denoted WRS_i, is computed as follows [2]:

$$WRS_i = r_i^{\frac{1}{u_i}} \tag{5}$$

where r is a uniformly selected real number in $[0, 1)$ and u_i is the document's uncertainty. We determine the WRS_i for each document and add to a minimum indexed priority queue. We then take k documents with the smallest WRS_i. These documents will typically have smaller certainty measures but it is still possible that a selected document has a higher certainty depending on the r generated. We believe this approach has a nice balance between preferring less certain documents and some exploration among certain documents. Another interpretation is that this selection method does not completely trust the classifier's predictions.

4.2.4 MoreCertainWRS. Analogous to *LessCertainWRS*, we compute a similar WRS_i but instead using certainty c_i:

$$WRS_i = r_i^{\frac{1}{c_i}} \tag{6}$$

4.2.5 Random. Finally as a baseline we randomly select k documents uniformly from U. Note that this method is equivalent to taking and annotating a simple random sample of the corpus as shown in Figure 3.

5 EXPERIMENTS AND RESULTS

This section presents the results of experiments using the presented selection methods. Based on Figure 3 we take 550 simple random samples from the corpus to create U_{INIT}. This is roughly 2% of the

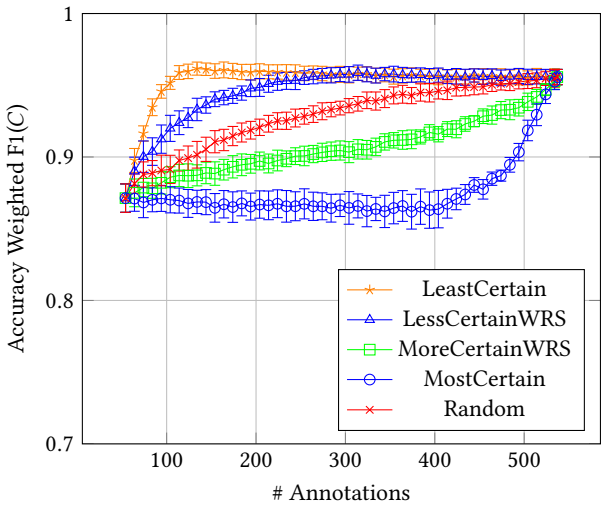

Figure 4: Annotations vs Corpus Accuracy with $|L_{init}| = 55$

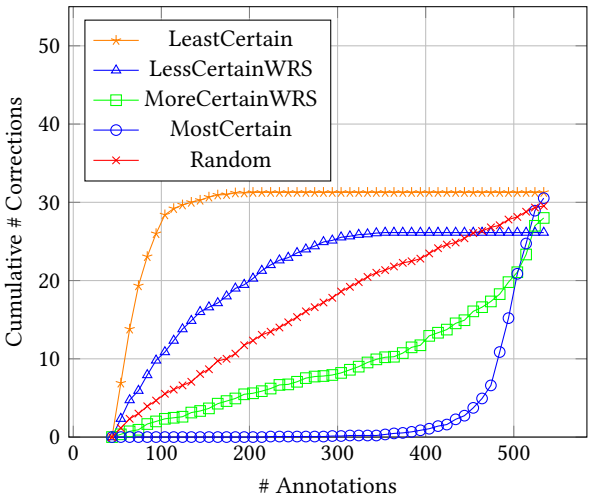

Figure 5: Annotations vs Corrections with $|L_{init}| = 55$

corpus size. We then take a simple random sample of 10% from U_{INIT} to create L_{INIT} (later we also look at taking 1% from U_{INIT}). This L_{INIT} of size 55 is unlikely to have all classes from Table 1 represented. We would expect that initial models perform poorly but improve as more samples are selected. For each experiment the weighted F1 of the model produced by the selection method at each iteration is evaluated on the corpus (the hold out of 26,336 documents). We repeat this experiment 20 times, take the average, and compute 95% confidence intervals.

5.1 Accuracy on Corpus Set

Figure 4 shows the results of the experiments. Note that the initial model and final model are the same for all selection methods (all methods eventually annotate the entire unlabelled set) and therefore are expected to produce the same accuracy at the first and last iteration.

We found the *MostCertain* method results to be interesting. We did not believe that this approach would do well but are surprised that it performs so poorly and then dramatically recovers between 400 to 500 annotations. We believe that because the selection method is only picking those documents it is most certain, it does not learn new information. The *MostCertain* method does appear to be mirrored by the *LeastCertain* method that very quickly performs well on the corpus. These results provide evidence that the certainty metric, and associated RF classifier predictions, are indeed useful estimates for selection.

The randomness in *MostCertainWRS* helps avoid the poor performance of *MostCertain* selection method, however, it does not clearly perform better than the benchmark *Random* method. Finally, *LessCertainWRS* performs well but not as good as *LeastCertain* taking until 250 annotations to produce a similar model. We thought the mix of less certain documents with more certain documents would result in the best method but clearly the *LeastCertain* method is the superior selection method.

5.2 Corrections

In typical situations, the accuracy of the selection method on the corpus as shown in Figure 4 is unknown. An annotator must use either the batch or remaining unlabelled documents to know when enough annotations have been made.

The number of times the user has to correct the proposed document type is important. Users have indicated that it is easier to confirm the proposed types versus having to correct them. At the first iteration all methods will have 0 corrections. Unlike the previous accuracy results, however, the final number of corrections may not be the same. We would expect *LeastCertain* to incur many corrections early and fewer later. Conversely, we would expect *MostCertain* to not have many corrections to begin and then more towards the end of the iterations.

Figure 5 indeed shows this to be the case. The figure shows the cumulative corrections versus the number of annotations from the same experiments conducted in Section 5.1. It is interesting to note the similarity of Figures 5 and 4. This suggests that the number of corrections can be used to infer the accuracy of the model. Assuming the entire unlabelled set U was to be annotated, it is also notable that *LessCertainWRS* requires fewer corrections than the other methods.

5.3 Average Unlabelled Certainty

Another attainable measure is the certainty of the remaining documents in U_i. Figure 6 shows the average certainty of the remaining unlabelled documents at each iteration with 95% confidence intervals.

Future work will consider using this measure or corrections as a stopping criteria. For the *LeastCertain*, *LessCertainWRS*, and *Random* methods it appears that when the average unlabelled certainty reaches about 0.8 these methods have achieved close to their maximum accuracy. The *MostCertain* and *MostCertainWRS* remove the most certain documents leaving U_i with less certainty and this is clearly shown in the figure. As the number of annotations increases, the size of $|U_i|$ gets smaller and we see more variability as shown

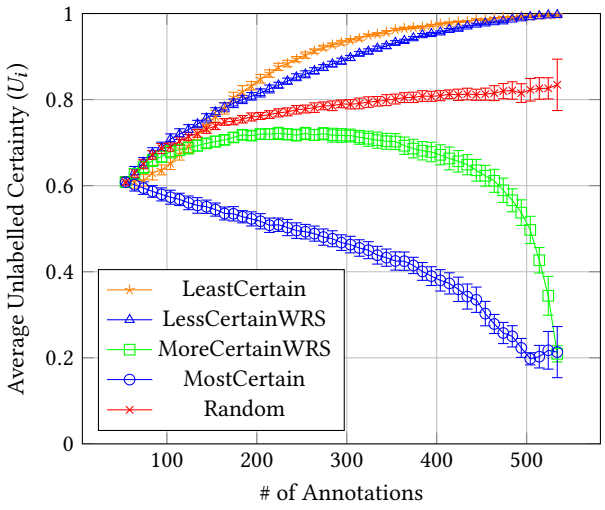

Figure 6: Annotations vs Unlabelled (U_i) Certainty

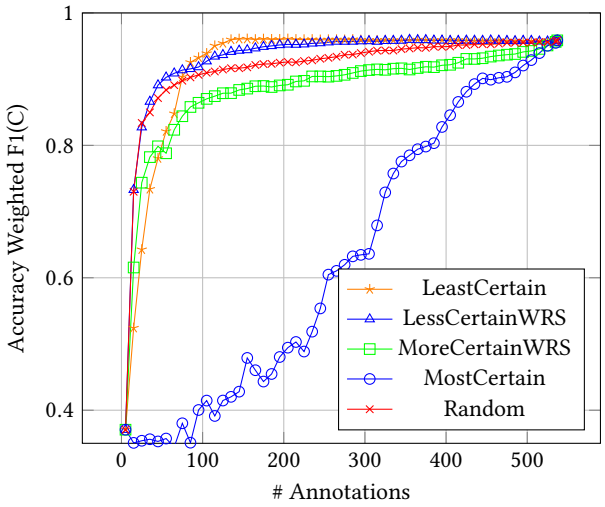

Figure 7: Annotations vs Corpus Accuracy with $|L_{init}| = 5$

by the confidence intervals after 500 annotations, particularly for *MostCertain*, *MostCertainWRS*, and *Random*.

5.4 Effect of Initial Labelled Set

The initial set of labelled documents L_{INIT} would certainty seem to affect the performance of the different approaches. We have shown that starting with a relatively good model (above 85% starting accuracy), the *LeastCertain* approach performs well.

We repeat the previous experiment 20 times but starting with a smaller initial labelled set of size $|L_{INIT}| = 5$ instead of $|L_{INIT}| = 55$. Given that there are ten document types this is a very under-specified model. Figure 7 shows the accuracy against the number of annotations (confidence intervals omitted for readability). The initial model's accuracy of 38% is not much better than the majority class classifier (per Table 1).

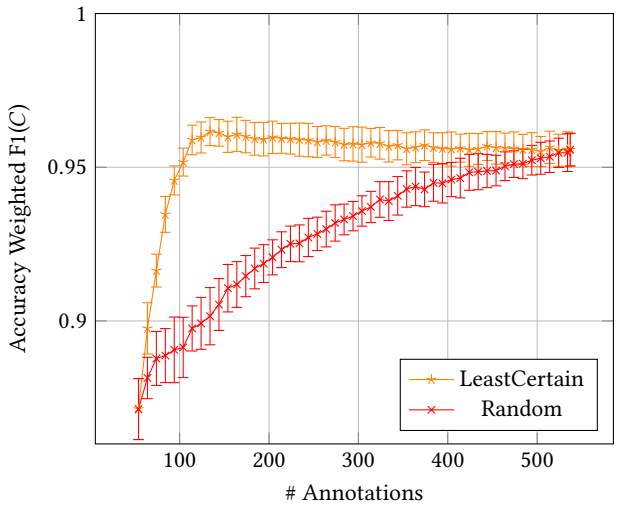

Figure 8: Random vs Least Certain Selection with $|L_{init}| = 55$

The *LeastCertain* method is greatly affected by the under-specified initial model and produces subsequent models with relatively poor accuracy until about 80 annotations. The methods with randomness (*LessCertainWRS*, *MoreCertainWRS*, and *Random*) are less affected with results comparable to Figure 4. The *LessCertainWRS* and *Random* methods are equivalent until approximately 40 annotations after which *LessCertainWRS* clearly performs better. If starting with a poor model, these results suggest that using the *LessCertainWRS* method is the best approach (possibly switching to the *LeastCertain* method at some point).

5.5 Random Sample vs Selection

Finally we revisit Figure 4 and restrict to just comparing the baseline *Random* method versus the *LeastCertain* selection method. Again, we start with an initial labelled set of size $|L_{init}| = 55$.

As previously indicated, the *LeastCertain* selection method consistently achieves a higher accuracy than the *Random* method for the same number of annotations. A significant result from Figure 8 is that if a user had done 400 annotations using simple random sampling they could have achieved similar accuracy with approximately only 100 annotations using the *LeastCertain* selection method. It is also notable that the *LeastCertain* selection method appears to slightly degrade from a high of 96.2% around 130 annotations to its final accuracy of 95.5% at 550 annotations.

6 CONCLUSION

This paper presented an active learning approach using uncertainty measures to assist in annotating a sample of documents. It demonstrated several selection methods' accuracy performance against the corpus along the model path. For our corpus and uncertainty measure the results confirm that adding more uncertain documents to the model more quickly leads to an accurate model. The paper also showed that the number of corrections has a strong relationship with the model's accuracy. Finally, the results show for a given

desired accuracy that annotating more uncertain documents first can reduce the annotation effort by as much as 75%.

7 ACKNOWLEDGMENTS

This work was supported in part by the University of Montevallo Contract #19-0501-001. The authors greatly appreciate the support of the airline company employees involved in the project. Without their efforts this research could not have been conducted.

REFERENCES

[1] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[2] Pavlos S. Efraimidis. 2015. *Weighted Random Sampling over Data Streams.* Springer International Publishing, Cham, 183–195. https://doi.org/10.1007/978-3-319-24024-4_12

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18. https://doi.org/10.1145/1656274.1656278

[4] Edwin Lughofer. 2012. Hybrid Active Learning for Reducing the Annotation Effort of Operators in Classification Systems. *Pattern Recogn.* 45, 2 (Feb. 2012), 884–896. https://doi.org/10.1016/j.patcog.2011.08.009

[5] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text Classification from Labeled and Unlabeled Documents Using EM. *Mach. Learn.* 39, 2–3 (May 2000), 103–134. https://doi.org/10.1023/A:1007692713085

[6] John C. Platt. 1999. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization.* MIT Press, Cambridge, MA, USA, 185–208.

[7] James Pope., Daniel Powers., J. A. (Jim) Connell., Milad Jasemi., David Taylor., and Xenofon Fafoutis. 2020. Supervised Machine Learning and Feature Selection for a Document Analysis Application. In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*. INSTICC, SciTePress, 415–424. https://doi.org/10.5220/0008925104150424

[8] Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola (Eds.). 1999. *Advances in Kernel Methods: Support Vector Learning.* MIT Press, Cambridge, MA, USA.

[9] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427 – 437. https://doi.org/10.1016/j.ipm.2009.03.002

[10] James V. Stone. 2013. *Information Theory: A Tutorial Introduction.* James V Stone.

[11] Meng Wang and Xian-Sheng Hua. 2011. Active Learning in Multimedia Annotation and Retrieval: A Survey. *ACM Trans. Intell. Syst. Technol.* 2, 2, Article 10 (Feb. 2011), 21 pages. https://doi.org/10.1145/1899412.1899414

[12] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2016. *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[13] Yiming Yang and Xin Liu. 1999. A Re-Examination of Text Categorization Methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Berkeley, California, USA) *(SIGIR '99)*. Association for Computing Machinery, New York, NY, USA, 42–49. https://doi.org/10.1145/312624.312647