Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.034157
*Article*

# Active Learning Strategies for Textual Dataset-Automatic Labelling

**Sher Muhammad Daudpota[1], Saif Hassan[1], Yazeed Alkhurayyif[2,*], Abdullah Saleh Alqahtani[3,4] and Muhammad Haris Aziz[5]**

[1]Department of Computer Science, Sukkur IBA University, Sukkur, 65200, Pakistan
[2]Al Quwayiyah College of Sciences and Humanities, Shaqra University, Shaqra, 15526, Saudi Arabia
[3]Self-Development Skills Department, Common First Year Deanship, King Saud University, Riyadh, 12373, Saudi Arabia
[4]STC's Artificial Intelligence Chair, Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh, 11451, Saudi Arabia
[5]College of Engineering & Technology, University of Sargodha, Sargodha, 40100, Pakistan

*Corresponding Author: Yazeed Alkhurayyif. Email: yalkhurayyif@su.edu.sa

Received: 07 July 2022; Accepted: 23 September 2022; Published: 30 August 2023

**Abstract:** The Internet revolution has resulted in abundant data from various sources, including social media, traditional media, etcetera. Although the availability of data is no longer an issue, data labelling for exploiting it in supervised machine learning is still an expensive process and involves tedious human efforts. The overall purpose of this study is to propose a strategy to automatically label the unlabeled textual data with the support of active learning in combination with deep learning. More specifically, this study assesses the performance of different active learning strategies in automatic labelling of the textual dataset at sentence and document levels. To achieve this objective, different experiments have been performed on the publicly available dataset. In first set of experiments, we randomly choose a subset of instances from training dataset and train a deep neural network to assess performance on test set. In the second set of experiments, we replace the random selection with different active learning strategies to choose a subset of the training dataset to train the same model and reassess its performance on test set. The experimental results suggest that different active learning strategies yield performance improvement of 7% on document level datasets and 3% on sentence level datasets for auto labelling.

**Keywords:** Active learning; automatic labelling; textual datasets

## 1 Introduction

Data is the oil of the 21st century-thanks to social media platforms, it is now abundantly available without much effort. Although an exponential increase in internet users and the phenomenal success of social media in the past couple of decades have resulted in the increased availability of huge amounts of unlabeled data, most natural language processing tasks require labelled data, especially supervised learning. Labelling data is still heavily reliant on human tagging efforts, which is expensive and tedious.

There have been some attempts at automatically labelling textual datasets without human involvement [1–3]; however, these approaches are very specific to one task and usually cannot be generalized. For example, some studies automatically use emoticons to learn sentiment in a tweet text. Therefore, it works fine with tweet text; however, it cannot be applied to movie reviews as they do not usually use emoticons like these are exploited within tweet text.

Active learning is a more generalized approach toward reducing manual labelling efforts on the dataset. Active learning is a fine balance between fully supervised learning (availability of labelled dataset) and unsupervised learning (absence of labelled dataset). Fig. 1 shows the basic model of a typical active learning strategy for labelling unlabeled datasets. It involves the following steps in the process:

a) An active learning dataset labelling process starts with an unlabeled dataset $\cup$ Which is easy to acquire as unlabeled and unstructured data is abundantly available these days.
b) A small $\mathcal{S} \subset \cup$ is selected based on some query strategy and forwarded to a human expert to label it.
c) Labelled instances are used to assess the performance of the selected machine or deep learning algorithms.
d) The process continues until performance is reached the desired level.

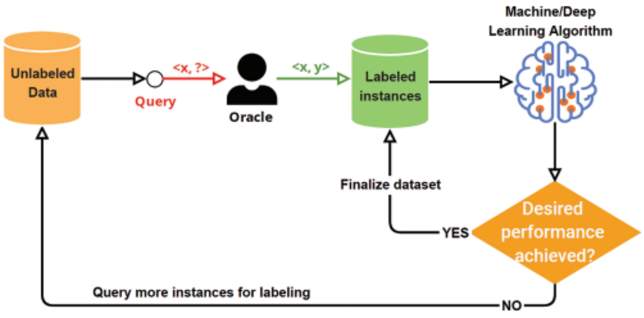
Unlabeled Data
<x, ?>
Query
Oracle
<x, y>
Labeled instances
Machine/Deep Learning Algorithm
Finalize dataset
YES
Desired performance achieved?
Query more instances for labeling
NO

**Figure 1:** Active learning basic model

The gist of an active learning model is that if an algorithm can choose which unlabeled instances to be labelled by the oracle, it will quickly start producing desired performance with less labelling efforts by a human.

Different strategies exist that an active learning algorithm might exploit to query an instance for oracle labelling. The simplest is a random selection, in which, starting with a small number of instances randomly, the model asks the oracle to label them initially and assesses the machine learning algorithm performance. In the subsequent passes, the algorithm randomly selects more unlabeled instances for oracle labelling until the machine learning algorithm reaches the desired performance level.

Random selection in active learning is not used in practice as it does not result in an advantage or realizing the actual strength of the active learning approach. In practice, the selection of subsequent instances for oracle labelling is either based on uncertainty or entropy. Uncertainty is learned by estimating the model's first choice label and second choice label probability differences. The lower the difference, the higher the uncertainty, so the model might want more uncertain instances labelled by the oracle.

On the contrary, entropy-based instances selection is based on how confident the model is about unlabeled instances. The more the entropy value, for instance, the higher the probability value for it to be selected for labelling by an oracle. The idea here is to boost the algorithm's learning where it is more confident and avoids any confusion for the algorithm.

Although the literature suggests many attempts at using active learning with different natural language processing tasks [4–7], there is a dire need to assess the performance of different query strategies in Natural Language Processing (NLP) tasks. More specifically, it is important to assess how different query strategies perform when instance size changes. For example, in sentiment analysis tasks, machine learning algorithms work at the sentence level, whereas in a typical document classification task, the instance size is much bigger and the algorithm has to work at a paragraph or even multiple paragraphs level.

### *1.1 Study Research Questions*

The main objective of this study is to assess the potential of active learning in automatic labelling of text datasets for different NLP tasks which use sentence level instances and document or paragraph(s) level instances. We have performed multiple experiments on different publicly available datasets to assess the performance and suitability of different active learning strategies. More specifically, the experiments are performed to answer the following research questions (RQs):

a) RQ1: Does active learning assist humans in the dataset labelling process?
b) RQ2: Is there any impact of selecting different active learning strategies on the automatic dataset labelling process?
c) RQ3: Which active learning strategies are better for long text (document level) and for short text (sentence level) in the textual dataset?

### *1.2 Study Research Contribution*

In line with our research questions, the following is the specific contribution of this study:

a) Identified active learning assistance in the manual dataset labelling process to reduce human efforts.
b) Assessed performance of different active learning strategies in dataset automatic labelling process.
c) Identified suitability of different active learning strategies for document level and sentence level textual datasets.

## 2 Related Work

The bottom line of an active learning algorithm is that human efforts in labelling data can be significantly reduced if the algorithm is allowed to choose instances that it wants humans to label. Suppose we have a pool $\mathcal{P}$ of unlabelled $n$ instances. Initially, the human might randomly select an $m$ number of instances where $\mathrm{m} \ll \mathrm{n}$. Based on m instances, a classifier is trained and an initial performance is obtained. Based on initial performance, an active learning algorithm initiates an iterative process involving humans to label new instances based on active learning strategies. The algorithm might select new instances based on what the algorithm decides. For example, it may choose those instances in which it is highly uncertain, or it might select those instances in which it is more confident. In an earlier case, the algorithm improves learning through the diversity of instances, in later, it increases confidence in similar instances. Nevertheless, how the algorithm decides for new instances to be labelled by a human or oracle is known as the query mechanism by the active learner.

## 2.1 *Active Learning Query Types*

There are three query types or mechanisms that active learners might exploit. All three query types have been discussed in detail below.

### *2.1.1 Pool-Based Sampling*

Fig. 2 shows the pool-based sampling process of querying samples for oracle labelling. The process starts with a learner, deep learning, or machine learning model. It estimates accuracy on initially randomly selected a small subset of instances. Based on its initial limited learning, it estimates probability scores for the remaining pool of unlabeled instances, which gives it information about the uncertainty of the whole unlabeled pool. Pool-based sampling has been applied in many active learning applications throughout the past two decades [8–11].

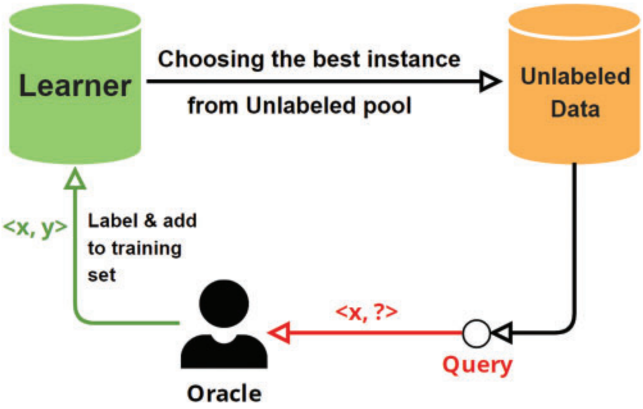

**Figure 2:** Pool-based sampling

### *2.1.2 Stream-based Selective Sampling*

In stream-based sampling, rather than considering a pool for informativeness, an individual sample was taken at a time [12,13]. It is queried from an unlabeled pool and the learner, based on the informativeness of the instance, decides whether to send the instance to an oracle for labelling or not. This process is computationally more expensive than pool-based sampling as the decision for labelling is taken in isolation of the pool.

### *2.1.3 Membership Query Synthesis*

Under this query mechanism, the learner, rather than picking an instance from the pool, generates one on its own for example, if it is about digits recognition, the learner might generate an image using augmentation on previous images to query oracle for its labelling [14]. This is usually a good approach when dealing with small datasets.

Regarding active learning for the class-imbalance problem, the study [15] used Active Learning to propose a solution for imbalanced class problems with the help of deep learning. It still lacks automatic labelling of the dataset, as it studies sequential active learning for balancing, which is performed as manual labelling, making it infeasible for batch mode active learning. Another study [16] improved performance for class-imbalance problems; however, humans do evaluations manually as it lacks an automatic model to address this problem. This study [17] proposed an active learning-based approach called Fair Active Learning (FAL) for balancing model accuracy and fairness by choosing instances to be labelled carefully. They worked on an unlabeled pool that increases the linear combination of misclassification error reduction and fairness improvement. Active learning is also used in a medical domain such as Heart Disease Prediction [18] along with a machine learning model for multi-label, in which five different strategies are used to reduce labelling costs by choosing the most relevant to be labelled to query.

### *2.2 Deep Learning*

Deep Learning is a part of machine learning, which is part of artificial intelligence. As such, deep learning is not a very new phenomenon in Artificial Intelligence (AI), as the building blocks of deep learning are neural networks that have existed in the field for many decades. Although in the past, neural networks were considered for supervised and unsupervised machine learning due to the requirement of neural networks for data and processing power, the machine learning community almost gave up on any further consideration of neural networks. It has only been recently reconsidered for machine learning tasks, thanks to the availability of abundant data as well as Graphical Processing Unit (GPU) based new architecture for computing. The biggest advantage of deep learning is its minimum requirement of human involvement in feature learning.

In 2011, Dan et al. started using deep learning in image classification with GPU-trained deep neural networks [19–21]; however, the watershed moment came in 2012 from the Hinton group, who submitted an entry in ImageNet image classification. Using traditional machine learning algorithms, the highest accuracy on the ImageNet dataset was about 74%. However, Krizhevsky et al. [22] used a convolutional neural network to bring accuracy to around 84% on 1000 ImageNet classes. Since this entry, deep learning has dominated the solution on ImageNet and till 2015, the accuracy crossed 95%, meaning the problem of ImageNet is now considered solved.

Although initially, deep learning made inroads in image and video processing, however, later in the years, due to improvements in recurrent neural network (RNN) and long short-term memory (LSTM) network, natural language processing tasks were also equally benefited. NLP tasks like sentiment analysis [23–28], document classification [29–34], topic modelling [35–38], seq2seq generation [39,40], etc., are now best suited to deep neural networks and their different variations.

The watershed moment in NLP with deep learning was observed with the introduction of the attention model in the deep neural network [41,42] and later the use of the attention model in transformer [43,44] development for different NLP tasks. Transformers have been extensively used in Bidirectional Encoder Representations from Transformers (BERT) models [45,46] as well as a new sensation in deep learning called Generative Pre-trained Transformer (GPT) [47,48] models from OpenAI. Both BERT and GPT models are now almost an automatic choice for many deep learning NLP tasks. Studies in [49] have used deep learning in applications like speech recognition and medical domains.

Literature suggests many uses of active learning with deep learning. Shen et al. [50] conducted an in-depth survey on using deep learning with active learning in the medical domain. The combination of deep Learning with active learning has also been used for named entity recognition [51]; however, to the best of our knowledge, there is no attempt to exploit deep learning in combination with active learning and explore its different strategies for automatic dataset labelling which is one of the major contributions of this study.

## 3 Methodology

Fig. 3 shows the complete methodology of the proposed approach to assess different query strategies' performance on the selected dataset.

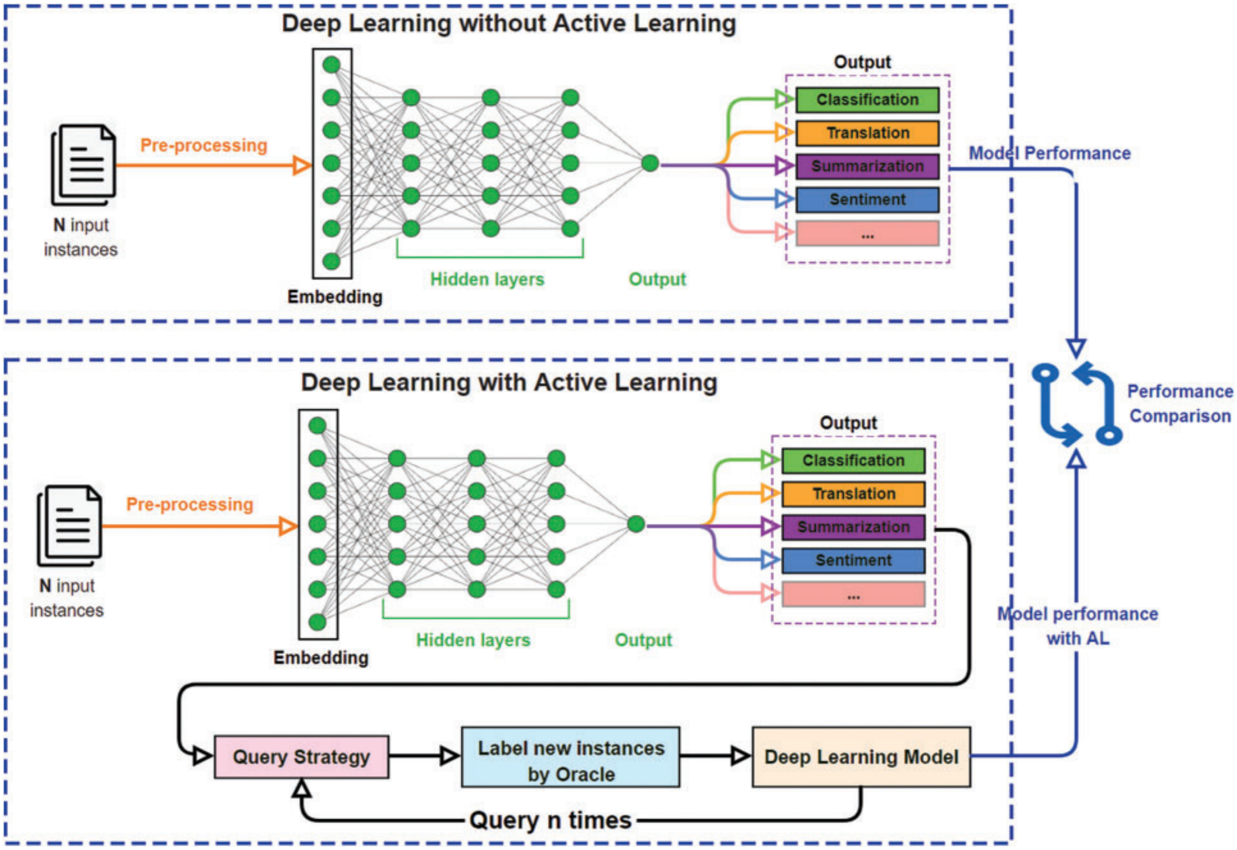

**Figure 3:** Methodology of the proposed approach

### 3.1 *Deep Learning Without Active Learning*

The sub-process starts with labelled instances from the selected dataset. In each case, we selected only a small subset of the original dataset ranging from 300 to 5000 instances to train the deep learning model. The actual strength of the deep learning model with the active learning model is in the small dataset; therefore, rather.

Using a full training dataset, we chose only a small subset of instances to assess the performance of the model and query strategies.

In the first phase, we selected all 300–5000 labelled instances for the different datasets and performed pre-processing and vectorization to provide input to the deep learning model. During pre-processing, we performed basic steps, including removal of stop words, punctuation, tokenization, conversion to lowercase, stemming and lemmatization.

After pre-processing, the next step is vectorization. Again, as our focus is not to achieve excellence in deep learning performance or beat previous benchmark results on the selected datasets, we employed simple one-hot encoding for converting words to vectors that can be fed to the deep learning model.

Fig. 4 summarises the deep learning model we used for classification purposes. Again, our objective was not to propose a state-of-the-art deep learning model or focus on its fine-tuning. We used a simple model with an input layer, a hidden layer and an output layer to perform classification and keep it the same with and without active learning approaches. The input neurons of the model are kept the same as the vocab size, which is kept at 10,000 to capture 10,000 top frequent words. The output layer has a variable number of neurons depending on the dataset's number of class labels.

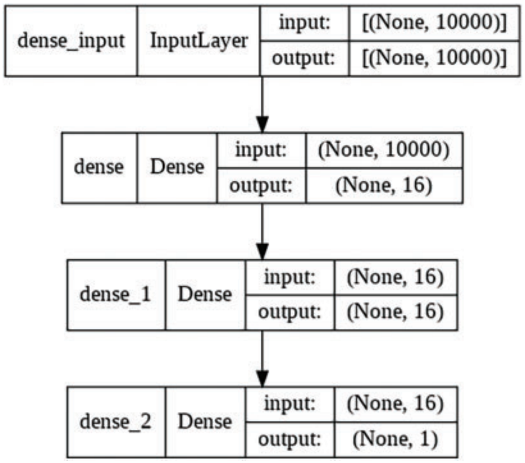

**Figure 4:** Deep learning model

### *3.2 Deep Learning with Active Learning*

This is the main crux of this study-here, we used the same number of 300–5000 instances for different datasets to assess the performance of the same classifier except involving different active learning strategies.

Instead of randomly selecting 300–5000 instances in this phase, we initially selected 50% of instances from the labelled dataset and maintained a pool of remaining training instances. The steps in this phase are mentioned below:

a) Train the model on initially selected 50% labelled instances.
b) Assess the model accuracy on the test set.
c) Use different active learning queries to select 10–25 instances from the training pool during each query to update the model's learning.
d) Re-train the model on initially selected 50% instances plus new instances selected through an active learning query strategy from the pool.
e) Assess the updated model's accuracy on the test dataset.
f) Continue the selection of new instances until you reach a total number of instances equal to 300–5000 for training for the different datasets.
g) Compare the performance of both models on the test dataset.

The main highlight of the proposed approach is step 3, where we queried the training pool to select instances for updating model learning. Here, unlike the traditional supervised approach, we provided the model with an opportunity to choose from the pool which instances it wants the oracle to label. In our case, we retrieved labels for those instances from the training set.

### *3.3 Active Learning Query Strategies*

As mentioned earlier, the main contribution of this study is to assess the impact of different active learning query strategies on classification performance. We summarize three popular query strategies in uncertainty sampling in active learning below.

#### *3.3.1 Classification Uncertainty*

Suppose a pool $\mathcal{P}$ of $n$ instances from which the model has to choose $m$ number of an instance where $m << n$, the pool is defined as

$$\mathcal{P} = \{\mathrm{x_1}, \mathrm{x_2}, \mathrm{x_3} \ldots \mathrm{x_n}\} \tag{1}$$

For an instance $\mathrm{x_i}$, the classification uncertainly is defined as,

$$\mathcal{U}(\mathrm{x_i}) = 1 - \mathrm{P}\,(\overline{\mathrm{x}}|\mathrm{x_i}) \tag{2}$$

where $\overline{\mathrm{x}}$ is the most likely prediction, for instance $\mathrm{x_i}$.

Suppose we have three class labels [A, B, C] and for an instance $\mathrm{x_i}$, the corresponding prediction probabilities are $[0.1, 0.7, 0.2]$. In this case, the most probable is label B for $\mathrm{x_i}$ and its corresponding $\mathcal{U}(\mathrm{x})$ is 0.3.

Table 1 shows a more detailed example of classification uncertainty. In this case, instance $\mathrm{x_2}$ will be selected as per the classification uncertainty strategy for oracle labelling.

**Table 1:** Example classification uncertainty

| Instance $(x_i)$ | P $(A)$ | $P\,(B)$ | $P\,(C)$ | $\mathcal{U}\,(x_i)$ |
|---|---|---|---|---|
| $x_1$ | 0.3 | 0.2 | 0.5 | 0.5 |
| $x_2$ | 0.4 | 0.3 | 0.3 | 0.6 |
| $x_3$ | 0.2 | 0.1 | 0.7 | 0.3 |

#### *3.3.2 Classification Margin*

Classification Margin for instance $\mathcal{M}\,(\mathrm{x_i})$ is defined as the difference between the probabilities of the most probable class label and the second most probable class label,

$$\mathcal{M}\,(x_i) = P\,(\overline{x_a}|x_i) - P\,(\overline{x_b}|x_i) \tag{3}$$

where $\overline{\mathrm{x_a}}$ is the class label, for instance, $\mathrm{x_i}$ with the highest probability and $\overline{\mathrm{x_b}}$ is the class label for the second highest probability. Table 2 shows a detailed example of the classification margin. In this case, again label $\mathrm{x_2}$ will be selected as the margin between the most probable and second most probable labels are lowest, which indicates the classifier is not confident about the class label.

#### *3.3.3 Classification Entropy*

Finally, the classification entropy for an instance $x$ with $n$ class probable class labels is defined as,

$$\mathcal{E}\,(x) = \sum_{i=1}^{n} P_i * log\,(P_i) \tag{4}$$

**Table 2:** Example classification margin

| Instance ($x_i$) | P ($A$) | $P$ ($B$) | $P$ ($C$) | $\mathcal{M}$ ($x_i$) |
|---|---|---|---|---|
| $x_1$ | 0.3 | 0.2 | 0.5 | 0.2 |
| $x_2$ | 0.4 | 0.3 | 0.3 | 0.1 |
| $x_3$ | 0.2 | 0.1 | 0.7 | 0.5 |

where $P_i$ is the probability, for instance, $x$ belonging to class label $i$. Table 3 shows a detailed example of classification entropy. In this example instance $x_1$ will be chosen for oracle labelling as it conveys the highest entropy.

**Table 3:** Example classification entropy

| Instance ($x_i$) | P ($A$) | $P$ ($B$) | $P$ ($C$) | $\mathcal{E}$ ($x_i$) |
|---|---|---|---|---|
| $x_1$ | 0.12 | 0.6 | 0.28 | 0.92 |
| $x_2$ | 0.03 | 0.9 | 0.07 | 0.38 |
| $x_3$ | 0.2 | 0.1 | 0.7 | 0.80 |

We label 10–25 instances during each query by querying the labelled dataset and train the model on an updated number of labelled instances. After 10−25 such queries, we assessed the model performance on the test dataset and compared it with the performance of the deep learning model without the involvement of active learning. There exist many implementations of active learning. However, we have used modAL: A modular active learning framework for Python[1].

## 4 Dataset

We carefully chose the dataset for our experiments keeping in view the following criteria:

a) The dataset should be publicly available to ensure our experiments are reproducible
b) A variety of datasets should be selected so that we assess the impact of sentence level instances as well as paragraph/document level instance
c) The dataset should be in the English language

Based on the above criteria, we selected four publicly available datasets that cover sentence-level as well as paragraph/document-level instances. Table 4 shows details of the selected datasets.

**Table 4:** Datasets selected for experiments

| Dataset | Level | Source | Source |
|---|---|---|---|
| Tweet emotions intensity | Sentence | [52] | Sentiment analysis |
| US airline tweet | Sentence | [53] | Sentiment analysis |
| BBC | Document | [54] | Document classification |
| BBC sports | Document | [54] | Document classification |

[1] https://modal-python.readthedocs.io/en/latest/

## 5 Results & Discussion on Research Questions

As shown in Fig. 3, we first applied deep learning without active learning, followed by active learning involvement with deep learning for each of our four selected datasets. As the power of active learning and its strategies is realized on a small dataset, we selected a small chunk of labelled instances (300–5000). During the first set of experiments, i.e., without the use of active learning, we fed all randomly selected labelled instances from the training pool to train the model, whereas, in the second set of experiments, we initially trained the same model on 50% labelled instances and allowed active learning model to choose to remain 50% instances from the pool using different strategies. Table 5 shows the experimental results on a subset of training instances selected from each training dataset pool.

**Table 5:** Experimental results-deep learning without and with active learning

| Dataset | Training | Test | Accuracy (DL) | Accuracy (DL+AL) | | |
|---|---|---|---|---|---|---|
| | | | | CU | CE | CM |
| Tweet emotions intensity | 5000 | 4211 | 0.54 | 0.52 | 0.52 | 0.53 |
| US airline tweet | 500 | 2928 | 0.73 | 0.73 | 0.76 | 0.76 |
| BBC | 500 | 445 | 0.91 | 0.95 | 0.96 | 0.94 |
| BBC sports | 300 | 148 | 0.90 | 0.97 | 0.96 | 0.97 |

### *5.1 Discussion*

*RQ1: Does active learning assist humans in the dataset labelling process?* It can be observed from the results that the involvement of active learning significantly reduces the efforts of labelling training instances, especially on document/paragraph level instances. It can be observed that with only 300 instances labelled by oracle for the BBC sports dataset, we achieved 97% accuracy in labelling the remaining 148 instances in the test dataset. Similarly, 445 test instances were labelled correctly for the BBC dataset with an accuracy of 96%. Although active learning involvement significantly improves performance when used with active learning, it is observed to perform even better when applied at the document level instead of the sentence level. The possible reason might be more information available for active learning strategies.

*RQ2: Does selecting different active learning strategies affect automatic dataset labelling?* Again, it can be observed from Table 5 that selecting different query strategies result in variation in accuracy on the test dataset. For example, classifier margin and uncertainty yield better accuracy on the BBC Sports dataset than classier uncertainly. Similarly, classifier margin performs better on the sentence level dataset than classifier entropy and uncertainty.

*RQ3: Which active learning strategies are better for long text (document level)? And which is better for short text (sentence level) in the textual dataset?* It can be observed that active learning improves the performance on the document level compared to the sentence level. Table 5 indicates that performance on BBC and BBC Sports datasets (document level datasets), the performance gain is about 7%, whereas, at the sentence level, it is up to 3%. In the Tweet Emotion Intensity dataset, a sentence-level dataset, the performance of active learning is even worse than the random selection of training instances.

## 6 Conclusion and Future Work

With the advent of social media, better storage capacity and high-speed internet, unstructured data in all forms, including text, audio and video, is abundantly available. However, we need to label it properly to use it for different supervised machines and deep learning tasks. Manual labelling is a tedious, time-consuming and expensive process. In most cases, labelling also requires expertise that is not easily available and expensive to exploit. In recent times, active learning has gained attention in auto-labelling instances with minimum efforts from oracles or humans. Active learning learns incrementally. It starts the process with a small set of labelled instances and its algorithm then decides which instances it wants for oracle labelling; thus, it is more like a human and algorithm cooperating for labelling the unlabeled data or more appealing human in loop terminology.

The results in this study suggest that by using active learning with deep learning, we can significantly reduce human efforts in the labelling process for the textual dataset. Results suggested that involvement of active learning improves labelling performance by up to 7% at document level datasets whereas up to 3% for sentence level datasets. It was observed that different strategies of active learning yield different accuracy on both sentence and document level datasets.

This work can be further extended by exploring other active learning strategies, including disagreement sampling, information density, etc. Furthermore, active learning with traditional machine learning algorithms can also be explored to assess the performance variation compared to deep learning. It would also be interesting to assess the model performance from a complexity and memory requirement perspective. Users of traditional machine learning algorithms are expected to require lesser memory and improved computational complexity. Therefore, experiments with traditional machine learning would add further insights to this study.

Active learning can also benefit from NLP tasks such as topic modelling and seq2seq generation. In topic modelling, active learning may assist in reducing the number of topics to reach a smaller subset of topics for a human to process for better accuracy. In seq2seq generation, active learning may assist in translation tasks from one sequence to another. For example, translation of English text to French or vice-versa. These exciting aspects of the study are yet to be explored.

**Acknowledgement:** The authors would like to thank the Deanship of Scientific Research at Shaqra University for supporting this work.

**Funding Statement:** The authors would like to thank the Deanship of Scientific Research at Shaqra University for supporting this work.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] A. Go, R. Bhayani and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, no. 12, pp. 2009, 2009.

[2] S. Fujimura, K. O. Fujimura and H. Okuda, "Blogosonomy: Autotagging any text using bloggers' knowledge," in *Proc. IEEE/WIC/ACM Int. Conf. on Web Intelligence (WI'07)*, Fremont, CA, USA, pp. 205–212, 2007.

[3] M. Enkhsaikhan, W. Liu, E. J. Holden and P. Duuring, "Auto-labelling entities in low-resource text: A geological case study," *Knowledge and Information Systems*, vol. 63, no. 3, pp. 695–715, 2021.

[4] F. Laws and H. Schütze, "Stopping criteria for active learning of named entity recognition," in *Proc. of the 22nd Int. Conf. on Computational Linguistics (Coling 2008)*, Manchester, UK, pp. 465–472, 2008.

[5] Y. Chen, S. Mani and H. Xu, "Applying active learning to assertion classification of concepts in clinical text," *Journal of Biomedical Informatics*, vol. 45, no. 2, pp. 265–272, 2012.

[6] V. C. Tran, N. T. Nguyen, H. Fujita, D. T. Hoang and D. Hwang, "A combination of active learning and self-learning for named entity recognition on Twitter using conditional random fields," *Knowledge-Based Systems*, vol. 132, pp. 179–187, 2017.

[7] S. das Bhattacharjee, A. Talukder and B. V. Balantrapu, "Active learning based news veracity detection with feature weighting and deep-shallow fusion," in *Proc. 2017 IEEE Int. Conf. on Big Data (Big Data)*, Boston, MA, USA, pp. 556–565, 2017.

[8] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine Learning Proceedings 1994*. New Brunswick, New Jersey, US: Elsevier, pp. 148–156, 1994.

[9] A. K. McCallumzy and K. Nigamy, "Employing EM and pool-based active learning for text classification," in *Proc. Int. Conf. on Machine Learning (ICML)*, San Francisco, CA, US, pp. 359–367, 1998.

[10] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proc. of the 2008 Conf. on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, pp. 1070–1079, 2008.

[11] A. G. Hauptmann, W. H. Lin, R. Yan, J. Yang and M. Y. Chen, "Extreme video retrieval: Joint maximization of human and computer performance," in *Proc. of the 14th ACM Int. Conf. on Multimedia*, Santa Barbara CA USA, pp. 385–394, 2006.

[12] L. E. Atlas, D. A. Cohn and R. E. Ladner, "Training connectionist networks with queries and selective sampling," *Advances in Neural Information Processing Systems*, vol. 16, pp. 566–573, 1990.

[13] A. Cohn, D. David, L. Atlas and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[14] D. Angluin, "Queries revisited," *Theoretical Computer Science*, vol. 313, no. 2, pp. 175–194, 2004.

[15] U. Aggarwal, A. Popescu and C. Hudelot, "Active learning for imbalanced datasets," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 1428–1437, 2020.

[16] C. Lin, M. Mausam and D. Weld, "Active learning with unbalanced classes & example-generated queries," in *Proc. of AAAI Conf. on Human Computation*, Zurich, Switzerland, pp. 98–107, 2018.

[17] H. Anahideh, A. Asudeh and S. Thirumuruganathan, "Fair active learning," *Expert Systems with Applications*, vol. 199, pp. 1–14, 2022.

[18] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri and H. Salem, "Multi-label active learning-based machine learning model for heart disease prediction," *Sensors*, vol. 22, no. 3, pp. 1–18, 2022.

[19] A. Ciresan, D. Claudiu, M. Ueli, M. Jonathan and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Twenty-second Int. Joint Conf. on Artificial Intelligence*, Barcelona, Catalonia, Spain, pp. 1237–1242, 2011.

[20] D. Ciregan, U. Meier and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, Rhode Island, USA, pp. 3642–3649, 2012.

[21] D. Ciresan, A. Giusti, L. Gambardella and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," *Advances in Neural Information Processing Systems*, vol. 25, no. 1, 2012.

[22] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 1, pp. 84–90, 2012.

[23] L. Zhang, S. Wang and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. e1253, 2018.

[24] Q. T. Ain, M. Ali, B. Hayat and A. Rehman, "Sentiment analysis using deep learning techniques: A review," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 424–433, 2017.

[25] A. Severyn and A. Moschitti, “Twitter sentiment analysis with deep convolutional neural networks,” in *Proc. of the 38th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Santiago Chile, pp. 959–962, 2015.

[26] L. M. Rojas-Barahona, “Deep learning for sentiment analysis,” *Language and Linguistics Compass*, vol. 10, no. 12, pp. 701–719, 2016.

[27] N. C. Dang, M. N. Moreno-Garc and F. La Prieta, “Sentiment analysis based on deep learning: A comparative study,” *Electronics*, vol. 9, no. 3, pp. 483, 2020.

[28] R. Batra, A. S. Imran, Z. Kastrati, A. Ghafoor and S. Shaikh, “Evaluating polarity trend amidst the coronavirus crisis in peoples’ attitudes toward the vaccination drive,” *Sustainability*, vol. 13, no. 10, pp. 5344, 2021.

[29] M. Z. Afzal, “Deepdocclassifier: Document classification with deep convolutional neural network,” in *2015 13th Int. Conf. on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisia, pp. 1111–1115, 2015.

[30] Z. Kastrati, A. S. Imran and S. Y. Yayilgan, “The impact of deep learning on document classification using semantically rich representations,” *Information Processing & Management*, vol. 56, no. 5, pp. 1618–1632, 2019.

[31] K. Kowsari, D. E. Brown, M. Heidarysafa and L. E. Barnes, “Hdltex: Hierarchical deep learning for text classification,” in *2017 16th IEEE Int. Conf. on Machine Learning and Applications (ICMLA)*, Cancun, Mexico, pp. 364–371, 2017.

[32] M. N. Asim, M. U. Ghani, M. A. Ibrahim and S. Ahmed, “Benchmarking performance of machine and deep learning-based methodologies for Urdu text document classification,” *Neural Computing and Applications*, vol. 33, no. 11, pp. 5437–5469, 2021.

[33] A. Adhikari, A. Ram, R. Tang and J. Lin, “DocBERT: Bert for document classification,” arXiv preprint arXiv:1904.08398, 2019.

[34] S. Shaikh, S. Daudpota, A. S. Imran and Z. Kastrati, “Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models,” *Applied Sciences*, vol. 11, no. 2, pp. 869, 2021.

[35] M. Jin, X. Luo, H. Zhu and H. H. Zhuo, “Combining deep learning and topic modeling for review understanding in context-aware recommendation,” in *Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, pp. 1605–1614, 2018.

[36] A. R. Pathak, M. Pandey and S. Rautaray, “Adaptive framework for deep learning based dynamic and temporal topic modeling from big data,” *Recent Patents on Engineering*, vol. 14, no. 3, pp. 394–402, 2020.

[37] Z. Cao, S. Li, Y. Liu, W. Li and H. Ji, “A novel neural topic model and its supervised extension,” in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 29, no. 1, Austin, Texas, USA, 2015.

[38] A. R. Pathak, M. Pandey and S. Rautaray, “Adaptive model for dynamic and temporal topic modeling from big data using deep learning architecture,” *International Journal of Intelligent Systems and Applications*, vol. 11, no. 6, pp. 13, 2019.

[39] P. Karpov, G. Godin and I. V Tetko, “A transformer model for retrosynthesis,” in *Int. Conf. on Artificial Neural Networks*, Munich, Germany, pp. 817–830, 2019.

[40] V. Kumar, A. Choudhary and E. Cho, “Data augmentation using pre-trained transformer models,” arXiv preprint arXiv:2003.02245, 2020.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, no. 1, pp. 1–5, 2015.

[42] J. Juraska and M. Walker, “Attention is indeed all you need: Semantically attention-guided decoding for data-to-text NLG,” arXiv preprint arXiv:2109.07043, 2021.

[43] I. V. Tetko, P. Karpov, R. Van and G. Godin, “State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis,” *Nature Communications*, vol. 11, no. 1, pp. 1–11, 2020.

[44] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue *et al.,* “Huggingface’s transformers: State-of-the-art natural language processing,” arXiv preprint arXiv:1910.03771, 2019.

[45] Y. Peng, S. Yan and Z. Lu, “Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets,” arXiv preprint arXiv:1906.05474, 2019.

[46] I. Tenney, D. Das and E. Pavlick, “BERT rediscovers the classical NLP pipeline,” arXiv preprint arXiv:1905.05950, 2019.

[47] L. Floridi and M. Chiriatti, “GPT-3: Its nature, scope, limits, and consequences,” *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020.

[48] Z. Hu, Y. Dong, K. Wang, K. W. Chang and Y. Sun, “Gpt-gnn: Generative pre-training of graph neural networks,” in *Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA, pp. 1857–1867, 2020.

[49] S. Budd, E. C. Robinson and B. Kainz, “A survey on active learning and human-in-the-loop deep learning for medical image analysis,” *Medical Image Analysis*, vol. 6, no. 1, pp. 102062, 2021.

[50] Y. Shen, H. Yun, Z. C. Liptonand and A. Anandkumar, “Deep active learning for named entity recognition,” arXiv preprint arXiv:1707.05928, 2017.

[51] S. Moon, S. Chung and S. Chi, “Bridge damage recognition from inspection reports using NER based on recurrent neural network with active learning,” *Journal of Performance of Constructed Facilities*, vol. 34, no. 6, pp. 4020119, 2020.

[52] S. M. Mohammad and M. Bravo, “Emotion intensities in tweets,” arXiv preprint arXiv:1708.03696, 2020.

[53] A. Rane and K. Anand, “Sentiment classification system of Twitter data for US airline service analysis,” in *2018 IEEE 42nd Annual Computer Software and Applications Conf. (COMPSAC)*, IEEE, Tokyo, Japan, vol. 1, pp. 769–773, 2018.

[54] D. Greene and P. Cunningham, “Practical solutions to the problem of diagonal dominance in kernel document clustering,” in *Proc. of the 23rd Int. Conf. on Machine Learning*, Pittsburgh, Pennsylvania, pp. 377–384, 2006.