

## RESEARCH ARTICLE

# Active Learning Based on Transfer Learning Techniques for Text Classification

**DANIELA ONITA** 

Department of Computer Science, University of Bucharest, 050663 Bucharest, Romania

Department of Computer Science, "1 Decembrie 1918" University of Alba Iulia, 515900 Alba Iulia, Romania

e-mail: daniela.onita@uab.ro

**ABSTRACT** Text preprocessing is a common task in machine learning applications that involves hand-labeling sets. Although automatic and semi-automatic annotation of text data is a growing field, researchers need to develop models that use resources as efficiently as possible for a learning task. The goal of this work was to learn faster with fewer resources. In this paper, the combination of active and transfer learning was examined with the purpose of developing an effective text categorization method. These two forms of learning have proven their efficiency and capacity to train correct models with substantially less training data. We considered three types of criteria for selecting training points: random selection, uncertainty sampling criterion and active transfer selection. Experimental evaluation was performed on five data sets from different domains. The findings of the experiments suggest that by combining active and transfer learning, the algorithm performs better with fewer labels than random selection of training points.

**INDEX TERMS** Active learning, active transfer learning, text classification, transfer learning.

## I. INTRODUCTION


Today, Machine learning techniques are used to solve problems in many domains. Some learning techniques require a large amount of data to provide a satisfactory result, such as deep learning. Even though the research area of automatic labelling is growing, there are many areas where manual labelling is essential. In many fields and applications, it is difficult to use unsupervised learning, and data labelling is still considered a difficult, expensive, and time-consuming task [4], [14], [24].

The goal of this study is to make the use of our resources in text categorization as efficient as possible. Automatic and semi-automatic annotation services facilitate the annotation process for text classification tasks [6], [7], [13], [17], but a real-data set improves the training process. Furthermore, a consistent set of annotated data is required to develop an annotation service. In numerous real-world applications, active learning (AL) and transfer learning (TL) have proved

their ability to build accurate models with a greatly reduced amount of training data [8], [19], [28], [32], [33].

In this study, a combination of AL and TL was investigated for effective learning under data sparsity conditions for text classification tasks. Five data sets from different domains were selected for experimental evaluation. A criterion for AL using data from different learning contexts was investigated, similar to how TL approaches work. The experimental protocol shows that the experiments were divided into three steps to demonstrate the ability of the proposed criteria. In the first step, we selected the learning algorithm that performs best for each data set. In the second step, we demonstrate the ability of the uncertainty sampling criterion compared to the random selection of training points, and then we demonstrate the efficiency of the proposed active transfer criteria.

The AL strategy that is proposed is similar to the one introduced in our preliminary work, which was presented in the European Symposium on Artificial Neural Networks in 2018 [20]. In our preliminary work [20], we examine the combination of AL and TL for image classification tasks. This work also focuses on making the best use of the available resources, but in this case, the text data sets were used for

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues .

the experiments. In the extended version of paper, we added experiments which demonstrate the benefits of the proposed method. The added contributions of the work described here compared to our preliminary work presented in [20] are:

- Examination of the combination of AL and TL for the text classification task. We demonstrate that the proposed criteria work without depending on the type of classification task.
- Experimental evaluation on multiple data sets. We extended the experimental evaluation by using data sets from different fields.
- Investigate several evaluation metrics to choose the classification algorithm. If in our preliminary paper, we only compared models using accuracy metrics, in this paper, we compared each algorithm using different types of performance measures: accuracy, precision, recall and F1 score.
- Demonstrate the performance of ATL criteria using another metric. To demonstrate the effectiveness of the proposed criteria, we examine the comparison of the precision metrics obtained with the random selection and with the two strategies for the active selection of training points: the uncertainty sampling criterion and the AT criterion. The performance of the proposed criteria is also proven in this case.
- Selecting the worst model for experiments. To demonstrate the performance of the AT criterion, we trained the data with the algorithm that gave the lowest performance.
- Comparison of the proposed strategies using deep learning algorithms. Even though deep learning algorithms need more data and the presented experiments used datasets with a small number of records, the performance of the ATL criterion is proven.

In the experimental evaluation, we studied the proposed active transfer (AT) criterion for five data sets from different domains. The performance of the five data sets is different, but the results show that uncertainty sampling selection performs better than random sampling. Moreover, when comparing the uncertainty sample selection with the AT criterion, the experiments show that the proposed method performs better than the uncertainty selection of training points for all data sets used in the experimental evaluation. Based on our experiments, we can state that the AT criterion performs better in different scenarios that we tested.

Methods for AL can be roughly divided into two categories: those with and without an explicitly defined objective function. The methods that were used in this paper, uncertainty sampling criterion and AT criterion are included in the last category.

The paper is organized as follows. This section finishes with related works. Section II, details the proposed approaches for efficiently training correct models with a greatly reduced amount of training data. Furthermore, the data sets are presented and the protocol used for experiments was described step by step. In Section III, we present the

experimental evaluation performed on the five data sets. In Section IV, we conclude and present directions for future research.

#### A. RELATED WORK

Some recent studies have examined the use of AL and TL in various learning tasks [10], [15], [27], [28], [29], [30]. The authors of [10] developed a deep learning-based method that targets low-resource environments for entity resolution through a novel combination of TL and AL. Its architecture allows the learning of a model to be transferred from a resource-intensive environment to a resource-poor one, and AL was used to select some informative examples to fine-tune the transferred model. In [15], the authors proposed a combination of active transfer learning and Natural Language Processing to improve liver volumetry using surrogate metrics with Deep Learning.

The authors of [27] also addressed the problem of learning with limited labeled data. They used a deep convolutional neural network to prove the performance of the active transfer learning for the image recognition tasks. The authors of [5] considered that the construction of an efficient deep neural network is largely needed on a large number of labeled samples that are available. This is the reason why they proposed a unified deep network, combined with active transfer learning, which can be trained using only minimally labeled training data for hyperspectral image classification. In [11] the authors have proposed a method based on active transfer learning and data augmentation for compensation of weak predictive power of neural networks on an unseen domain.

In [28], the authors proposed a combination of TL and AL to create a learning system to detect atrial fibrillation. Their goal is the same as ours: To reduce data and cost and to develop a more cost-effective solution to their task.

The authors of [31] present a method that uses active learning to minimize the need to annotate the majority of examples in the data set. Their goal is to develop CAD systems by simplifying the tedious task of labeling images while maintaining similar performance to state-of-the-art methods. In [18], the authors combine transfer learning, active learning, and pseudo-labeling to develop pseudo-active transfer learning. Their method translates the complex description of intrusion alarms for analysts with confidence.

In [2], it is proposed as an alternative to the conventional criterion in AL that actively selects questions by using available preference data from other users in a preference learning scenario. The authors of [20] explore the combination of AL and TL for efficient learning in data scarcity for image classification. Experimental results suggest that by combining AL and TL on a target domain, they can learn faster with fewer labels than with random selection. Similarly, the goal of this work is to use the available resources as efficiently as possible, but in this case, text data sets were used for the experiments. In [20], we applied the same criteria for image classification.

## II. MATERIALS AND METHODS

The following methods are already explained in our preliminary paper [20].

### A. ACTIVE TRANSFER LEARNING

Active Learning [26] refers to a group of approaches that may be used when labeling points is difficult, time-consuming, and costly. The theory underlying AL is that by selecting training sites optimally rather than randomly, improved performance can be achieved.

There is a trend to improve the performance of AL methods by combining them with heuristics designed either for the context in which they are applied or by the models they use, such as using unlabeled data, exploiting clusters in the data, diversifying the set of hypotheses, or adapting the AL to other learning techniques such as Gaussian processes.

#### 1) UNCERTAINTY SAMPLING CRITERION

The uncertainty sampling criteria [26] is an AL approach in which an active learner labels the example for which the model's predictions are the most uncertain. The prediction's uncertainty can be assessed, for example, using Shannon's entropy

$$\text{Uncertainty}(x) = - \sum_y p(y|x) \log p(y|x). \quad (1)$$

where  $x$  is the point to be labeled and  $y$  denotes a possible label for  $x$ . This technique simplifies to querying points with prediction probability close to 0.5 for a binary classifier. Intuitively, this technique seeks to discover the decision boundary as quickly as feasible, as suggested by the regions where the model is most unclear.

#### 2) ACTIVE TRANSFER CRITERION

Transfer learning [22] is a technique for transferring knowledge from one activity to another. It is influenced by psychological research on learning transfer, notably the dependence of human learning on past experience. Because the psychological theory of transfer of learning assumes task similarity, TL algorithms are utilized when training data for the target task is comparable but not identical to that of the source task. TL could be performed in the context of learning algorithms by transferring learned features and parameters from one algorithm to another.

We propose here an AL criteria we term AT, which is specially designed to use the AL and TL parameters. The fundamental idea behind the AT criteria is to employ learning with many data sets to determine knowledge obtained with a new data point by using the learned models of earlier data sets.

For the prediction probability related to alternative models, we shall use the notation shown below

$$p_m(y|x) \equiv p(y|x, M_m). \quad (2)$$

where  $M_1, \dots, M_M$  denotes the data set unique to each task. Inspired by [16], we calculate the average Kullback-Leibler

(KL) divergence of individual forecasts from the average:

$$\text{AT}(x) = \sum_{m=1}^M \frac{1}{M} \text{KL}[\bar{p}(\cdot|x) \| p_m(\cdot|x)], \quad (3)$$

with  $\bar{p}(\cdot|x)$  the average predictive probability of the entire committee.

For discrete probabilities, the KL divergence is defined as

$$\text{KL}[p_1(\cdot|x) \| p_2(\cdot|x)] = \sum_c p_1(y|x) \log \frac{p_1(y|x)}{p_2(y|x)}. \quad (4)$$

The KL divergence may be thought of as a distance between probabilities, where we misused the concept of distance since the KL-divergence is not symmetric, i.e.,  $\text{KL}[p_1 \| p_2] \neq \text{KL}[p_2 \| p_1]$ . This disadvantage of the KL-divergence can be avoided by using a symmetric measure, such as  $\text{KL}[p_1 \| p_2] + \text{KL}[p_2 \| p_1]$ . In [16], the disagreement is computed between committee members constructed based on the current model, i.e., the committee changes with every update and the criterion has to be recomputed with every update. A committee of models learned on different tasks is fixed and thus selecting examples solely based on it leads to a fixed instead of an active design: all examples can be ranked beforehand [2].

### B. DATA SETS AND DATA PRE-PROCESSING

Five data sets were chosen for the following experiments. For each data set a binary classifier was considered.

**IMDB data set** [12] is a public data set for binary sentiment classification. The IMDB data set contains 50K movie reviews written in English that is used for natural language processing or text analytics.

**Spam data set** [1] is a collection of SMS-tagged messages that have been collected for SMS Spam research. The data set consists of 5,574 messages that are written in English, and tagged according to being ham (legitimate) or spam.

**Amazon product review data set** [9] consists of a few million Amazon customer reviews and star ratings. In the rating column, the rating is from 1 to 5. A new column was added for binary sentiment analysis: ratings 1 and 2 were converted as bad reviews, 4 and 5 as good reviews, and 3 were eliminated.

**PadChest data set** [3] is a public corpus that was collected in Spain at Hospital San Juan from 2009 to 2017. It includes more than 160k X-rays images. Radiologists assessed the X-ray pictures, and each image was paired with a report written in Spanish. The remaining reports were labeled using a supervised technique based on a recurrent neural network with attention mechanisms, with 27 percent manually annotated by trained clinicians. In the experimental evaluation, it was selected only the radiologists' reports. These reports were used for binary text classification where the label of each report was normal or anomaly [21].

**Movie Review Polarity data set** [23] consists of 2000 documents about movie reviews. The data set contains 1000 positive reviews about a movie and 1000 negative reviews.

For each data set, the following data preprocessing techniques were performed: Deletion of all null values, if any, and extraction of features from text files. In order to run machine learning algorithms for text classification, the text files must be converted into numerical feature vectors. A bag-of-words model was used. The 'CountVectorizer' from the scikit-learn library [25] was used to segment each text file into words. The 'CountVectorizer' develops a vector of all the words in the string. To develop the vector, it needs to count how many times each word occurs in each document, and finally, assign an integer ID to each word. The returned document is a document term matrix of size [n-samples X n-features].

### C. EXPERIMENTAL PROTOCOL

To achieve the goal of the paper, the first step was to compare different machine learning algorithms for each data set. To measure the performance of each machine learning algorithm, different types of performance measures were used: accuracy, precision, recall and F1 score. The results are presented in Section III. The best algorithm was used for active and random selection, and for storing a pre-trained model for each data set. For three data sets, we considered about 5,000 records, except for the Spam data set and the Movie Review Polarity data set, which are smaller. For the Spam data set, we considered about 4,500 records for the target set. For the Movie Review Polarity corpus, we considered about 1,500 records for the target set.

We then compared the active selection of training points with the random selection of training points. The training data was used as a pool from which points were randomly or actively selected for labeling. Once a point was selected, either actively or randomly, it was added to the training data and removed from the unlabeled data. With the updated training data set, the model was re-trained and predictions were made using the validation set. Results were averaged over 20 data splits into training, unlabeled, and validation sets. All algorithms were trained with 50 randomly selected and active data points.

The third step is to compare the active selection of training points to AT selection. Since each data set has a binary classifier, the point for which the prediction probabilities are closest to 0.5 was selected for the uncertainty sampling criterion. For the AT criterion, a portion of each data set is selected as the target. Each of the three data sets contains about 5,000 records, except for the Spam data set which consists of about 4,500 records, and the Movie Review Polarity corpus, for which 1,500 records were considered. The target data set will be trained on a data set created from the rest of the data set using a pre-trained model. To obtain the pre-trained models for each data set, we trained the data using the algorithm that performed best in step one of the experiments. The point chosen was the one with which the other model disagreed the most.

To demonstrate the performance of the AT criterion, we trained the data with the algorithm that yielded the lowest

performance. The following algorithms were considered: Decision Tree Classifier for the IMDB data set, GaussianNB for the Spam data set, Logistic Regression for the Amazon data set, Decision Tree for PadChest, and KNeighborsClassifier for the Movie Review Polarity data set. For each data set, we compared the three criteria for training point selection: random, active, using the uncertainty sampling criterion, and using the proposed AT criterion. The rest of the experimental protocol remained the same.

One more proof that AT criterion performs better than the uncertainty sampling criterion is that we trained the data using a deep learning model. We used a Sequential model with one hidden layer and 2.581 parameters. For the hidden layer, we used a rectified linear activation function and the output layer was activated by a Sigmoid function that is frequently used for classification. It is important to mention that we tried different architectures for deep learning models, but increasing the number of hidden layers does not mean that the performance of the model is increased too.

### III. RESULTS

In the first step, the following algorithms were compared: Logistic Regression, Linear Support Vector Classification, Decision Tree Classifier, Random Forest Classifier, Gaussian Naive Bayes Classifier, Multilayer Perceptron Classifier, K-nearest neighbors Classifier. We used accuracy (mean  $\pm$  standard deviation), precision, recall and F1-score as measures of performance.

Table 1 shows the comparison of the different machine learning algorithms for the IMDB data set. The last column of the table shows which algorithm provides the best score for each performance measure of performance. For the IMDB data set, the Random Forest Classifier provided the best accuracy and precision and the KNeighbors Classifier provided the best recall and F1 score. Based on the obtained scores, the Random Forest Classifier was used as the classifier for the IMDB data set.

Table 2 shows the results of applying different machine learning algorithms for the Spam data set. The best result was obtained with the Random Forest classifier for all measures that were compared for model performance.

Table 3 shows the comparison of seven machine learning classifiers for the Amazon product review data set. The KNeighbors classifier was used in the experimental evaluation for the Amazon data set because this classifier gave the best results in terms of accuracy, recall and F1 score.

As shown in Table 4, when comparing the learning algorithms for the PadChest data set, the best accuracy and precision were obtained with the Random Forest Classifier. For this reason, the Random Forest Classifier was used as the learning algorithm in the experiments with the PadChest data set.

Table 5 shows the comparison of seven machine learning algorithms for the Movie Review Polarity data set. For this data set, the MLP Classifier provides the best score for all performance measures.

**TABLE 1. Comparison of different learning algorithms for IMDB data set.**

	Logistic Regression	Support Vector Classifier	Decision Tree	Random Forest	Gaussian Naive Bayes	MLP Classifier	KNeighbors Classifier	Best score
Accuracy	0.5318	0.53172	0.5088	0.5362	0.5242	0.5114	0.51016	Random Forest
Precision	0.531619	0.531535	0.508909	0.538991	0.524671	0.511526	0.508032	Random Forest
Recall	0.53496	0.53496	0.5076	0.49928	0.52096	0.50512	0.642	KNeighbors Classifier
F1 Score	0.533272	0.533231	0.508186	0.518354	0.522516	0.508163	0.567174	KNeighbors Classifier

**TABLE 2. Comparison of different learning algorithms for Spam data set.**

	Logistic Regression	Support Vector Classifier	Decision Tree	Random Forest	Gaussian Naive Bayes	MLP Classifier	KNeighbors Classifier	Best score
Accuracy	0.733092	0.725066	0.866724	0.866724	0.66388	0.793129	0.861368	Random Forest
Precision	0.738172	0.730122	0.849169	0.849169	0.684567	0.832956	0.843427	Random Forest
Recall	0.778523	0.778523	0.92349	0.92349	0.778523	0.714246	0.922148	Random Forest
F1 Score	0.757164	0.752774	0.880476	0.880476	0.712222	0.756016	0.876555	Random Forest

**TABLE 3. Comparison of different learning algorithms for Amazon product review data set.**

	Logistic Regression	Support Vector Classifier	Decision Tree	Random Forest	Gaussian Naive Bayes	MLP Classifier	KNeighbors Classifier	Best score
Accuracy	0.6966	0.7342	0.7508	0.7508	0.6932	0.697	0.751	KNeighbors Classifier
Precision	0.817775	0.808147	0.807387	0.807387	0.840517	0.816733	0.807329961	Gaussian Naive Bayes
Recall	0.829798	0.887374	0.912374	0.912374	0.812121	0.830556	0.912878788	KNeighbors Classifier
F1 Score	0.758323	0.823517	0.84362	0.84362	0.734222	0.759399	0.843924811	KNeighbors Classifier

**TABLE 4. Comparison of different learning algorithms for PadChest data set.**

	Logistic Regression	Support Vector Classifier	Decision Tree	Random Forest	Gaussian Naive Bayes	MLP Classifier	KNeighbors Classifier	Best score
Accuracy	0.6818	0.6838	0.623	0.6924	0.3784	0.631	0.6326	Random Forest
Precision	0.426369	0.350649	0.367437	0.536775	0.308953	0.35754	0.343183	Random Forest
Recall	0.01219	0.003846	0.289917	0.102628	0.820507	0.228339	0.192425	Gaussian Naive Bayes
F1 Score	0.023326	0.007513	0.324071	0.171905	0.416197	0.27447	0.245449	Gaussian Naive Bayes

As described in Section II-C, after selecting the best algorithm for each data set, the next step is to use this algorithm to compare the active selection for training and the random selection.

For three of the data sets, we selected 5,000 records, except for the Spam data set and the Movie Review Polarity corpus, which are smaller. Each data set was divided into a

training data set, an unlabeled data set, and a validation data set. The training data set was used as a pool from which points were randomly or actively selected for labeling. For each type of selection, a point was selected, then added to the training set and deleted from the unlabeled data set. The new training set was used to train the model and the validation set was used to make predictions. Uncertainty



TABLE 5. Comparison of different learning algorithms for movie review polarity data set.

	Logistic Regression	Support Vector Classifier	Decision Tree	Random Forest	Gaussian Naive Bayes	MLP Classifier	KNeighbors Classifier	Best score
Accuracy	0.807499	0.788	0.629	0.8	0.728	0.8240	0.590	MLP Classifier
Precision	0.81088	0.7876	0.6276	0.8009	0.7741	0.8240	0.6642	MLP Classifier
Recall	0.8019	0.789	0.637	0.799	0.645	0.8240	0.377	MLP Classifier
F1 Score	0.8063	0.7882	0.6316	0.7997	0.7029	0.8237	0.4730	MLP Classifier

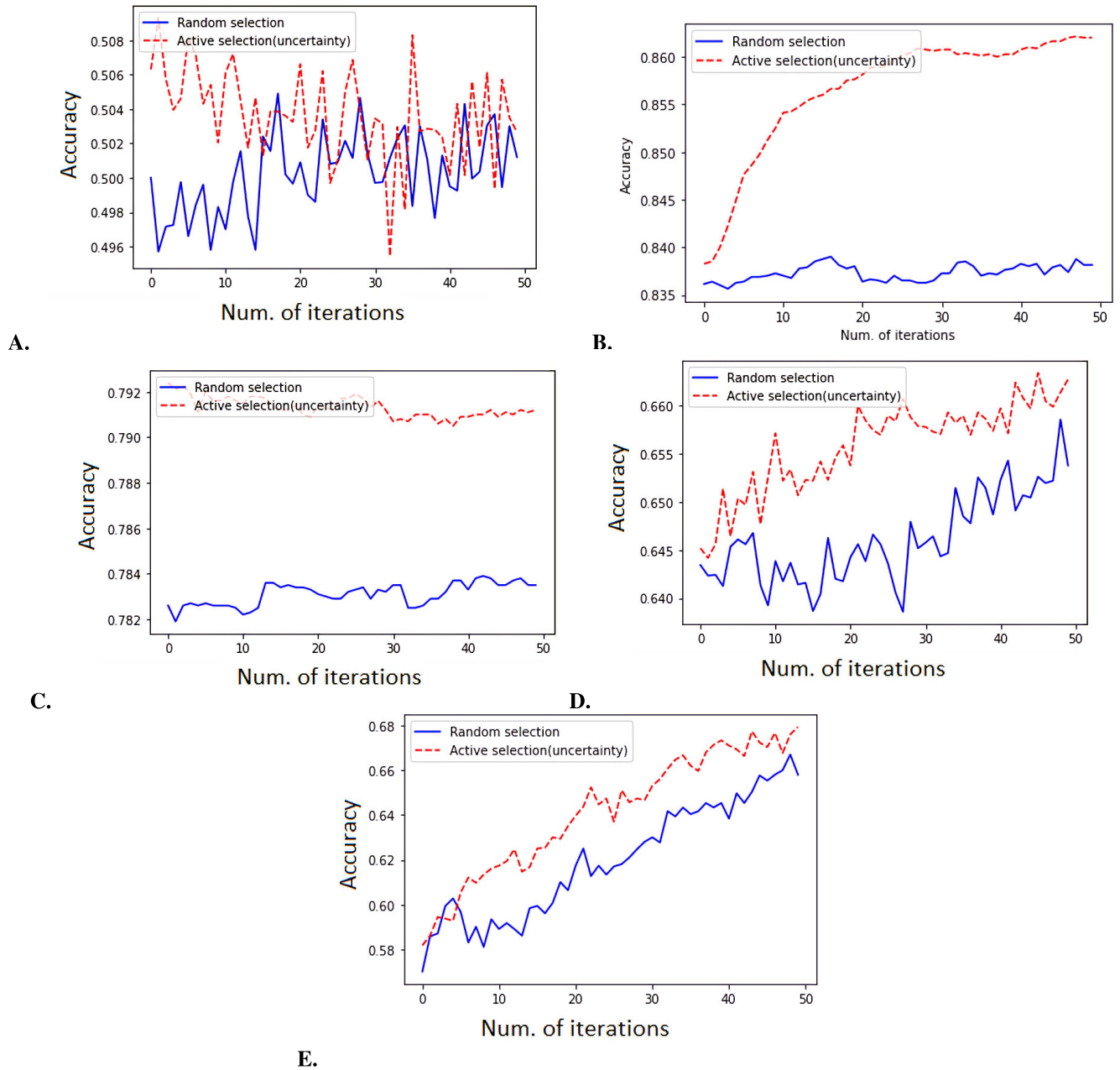
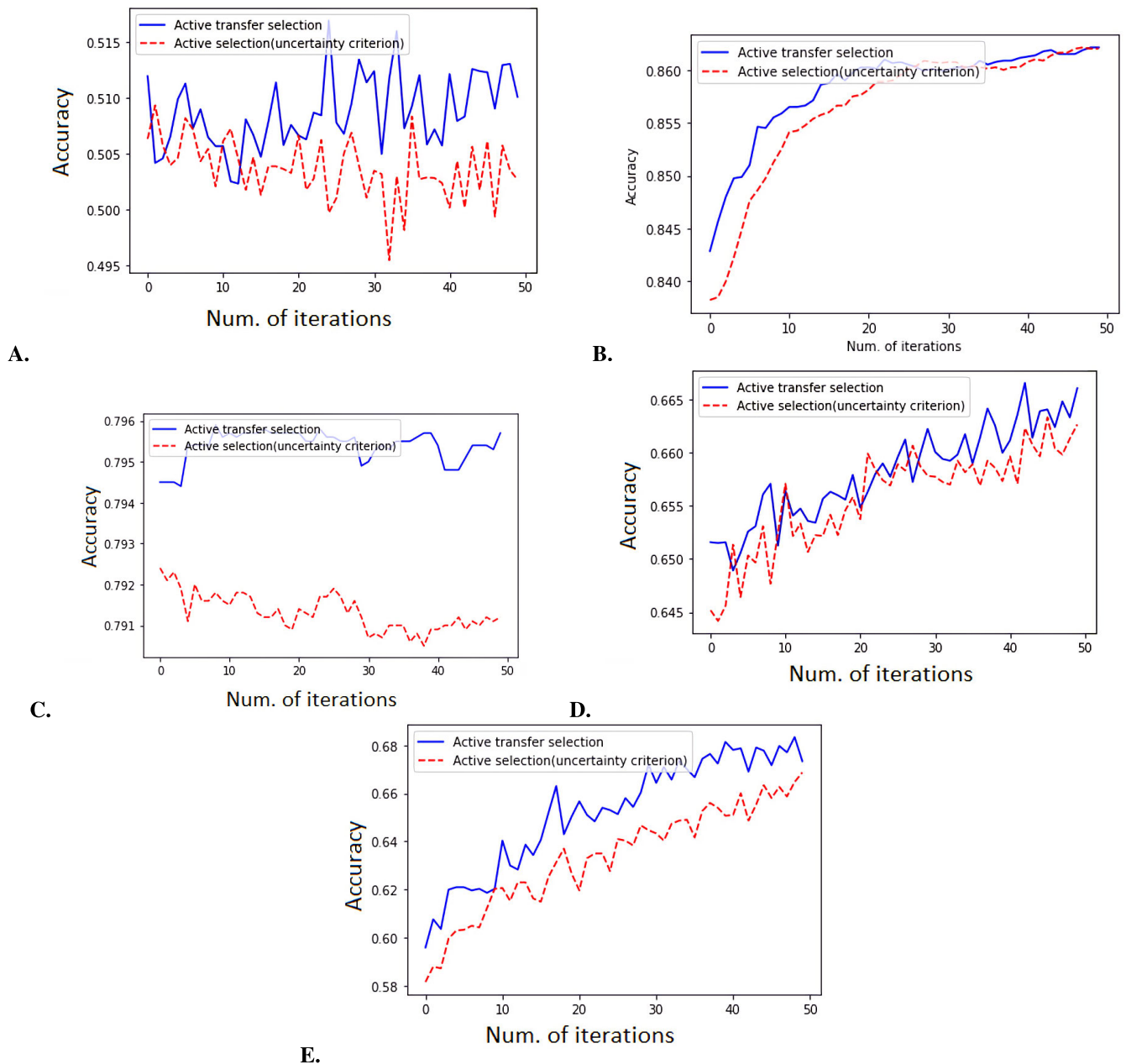


FIGURE 1. Comparison of accuracy (mean  $\pm$  standard deviation) derived from random versus active training point selection. (A.) IMDB data set. (B.) Spam data set. (C.) Amazon product review data set. (D.) PadChest data set. (E.) Movie review polarity data set.

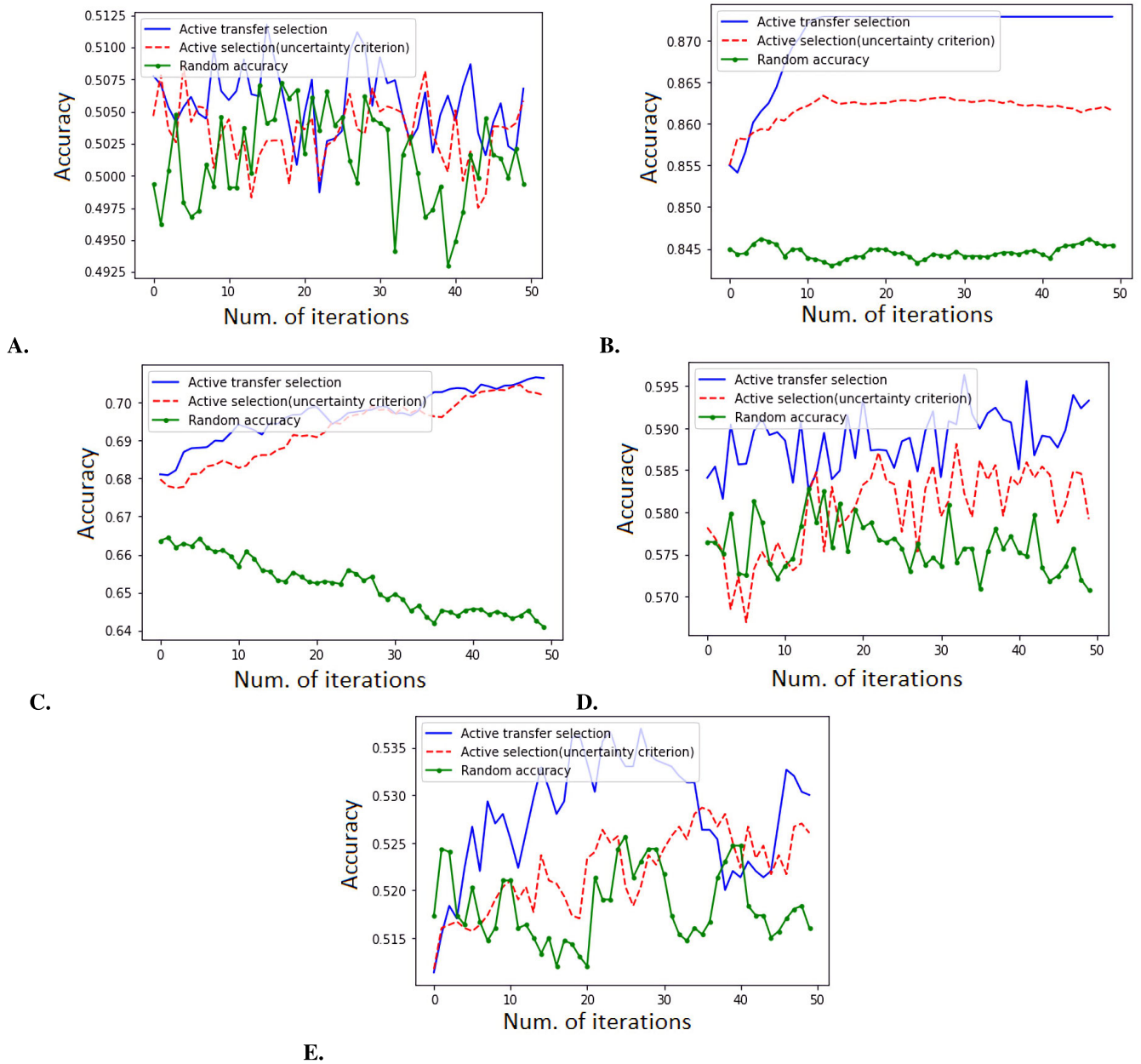


**FIGURE 2.** Comparison of accuracy (mean  $\pm$  standard deviation) obtained with two strategies of actively selecting training points. (A.) IMDB data set. (B.) Spam data set. (C.) Amazon product review data set. (D.) PadChest data set. (E.) Movie review polarity data set.

sampling criteria were used for the active selection of training sites.

Figure 1 shows the comparison of the accuracy obtained when training points were randomly and actively selected using the uncertainty sampling criterion for each data set used in the experimental evaluation. For each data set, the results obtained by active selection outperform those obtained by random selection. In the plot of the Spam data set (Figure 1 B.) and the Amazon data set (Figure 1 C.), the difference between the accuracy obtained by active selection of training points and random selection is more evident.

After comparing the random selection of training points with the active selection applied using the uncertainty sampling criterion described in Section II-A1, the next objective was to compare the active selection with the AT selection of training points. As mentioned earlier, the uncertainty sampling criterion was used for the active selection of training points. Since each data set has a binary classifier, the prediction probabilities closest to 0.5 were selected to calculate the uncertainty sampling criterion. For the AT criterion, we formed a target data set containing about 5,000 records for three of the data sets, except for the Spam



**FIGURE 3.** Comparison of accuracy obtained using the algorithm which returns lower performance. The training points were selected randomly and using the two strategies of actively selecting training points: uncertainty sampling criterion and AT criterion. (A.) IMDB data set. (B.) Spam data set. (C.) Amazon product review data set. (D.) PadChest data set. (E.) Movie review polarity data set.

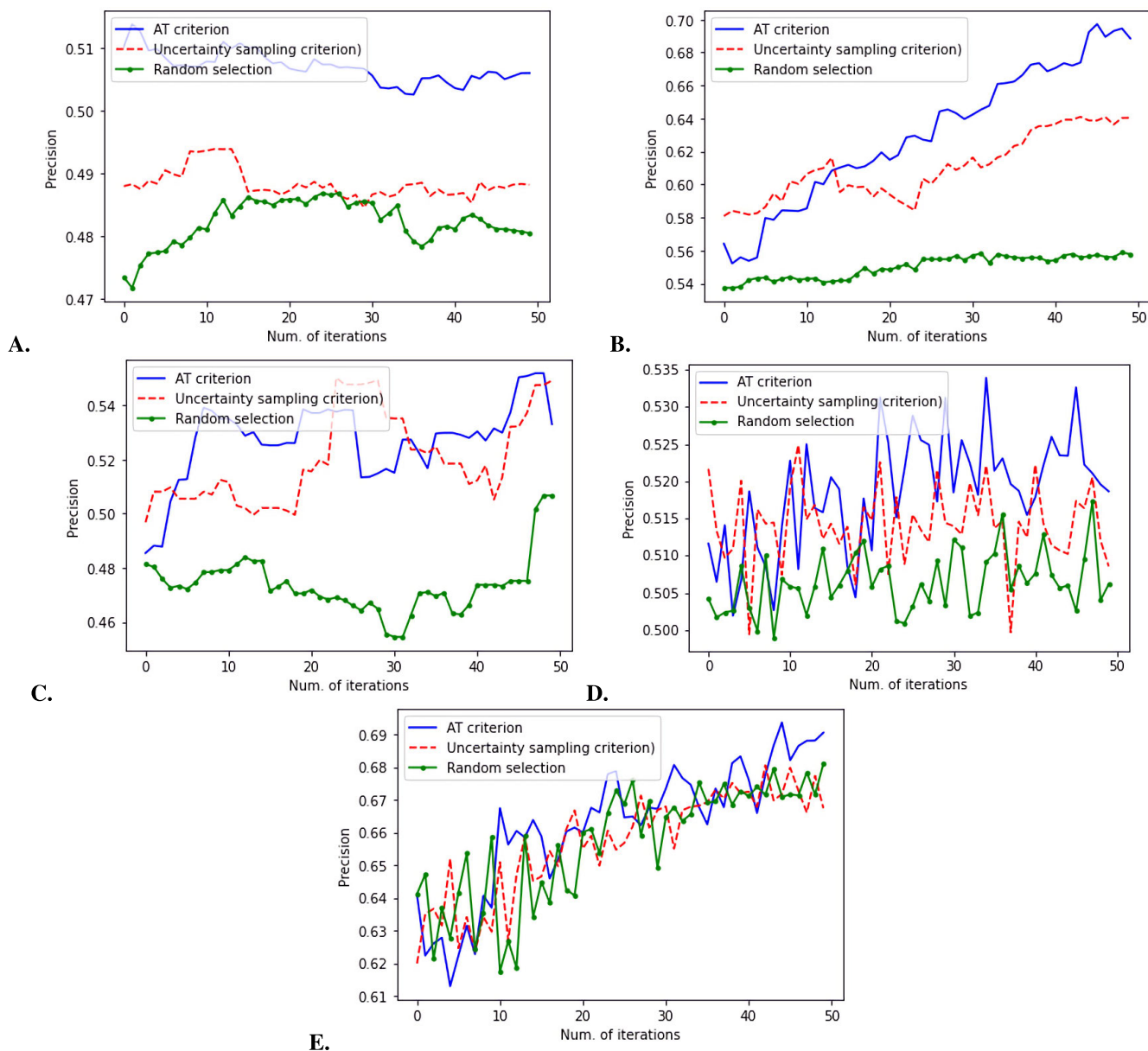
and Movie Review Polarity data sets, which are smaller. The target data set of the Spam data set contains only 4,500 records and for the Movie Review Polarity data set the target data set consists of 1,500 records. The rest of the records from the data sets were used to obtain the pre-trained models. When calculating the criterion AT, the point where the other model deviates the most was selected. The selected point was added to the training data set while it was removed from the unlabeled data set.

The plots in Figure 2 compare the accuracy achieved with AL using two criteria for selecting active points: the

uncertainty sampling criterion and the AT criterion. The plots show that the AT criterion performs better than the AL criterion using the uncertainty sampling criterion for each data set used in the experimental evaluation. For three of the five data sets, the accuracies achieved are close, but for the Spam data set (Figure 2 B.) and for the Amazon product review data set (Figure 2 C.), the difference between the two types of AL criteria is more pronounced.

It is important to note that the proposed AT criterion performs better for all data sets, even though the Spam and the Movie Review Polarity data sets are smaller. For the Spam





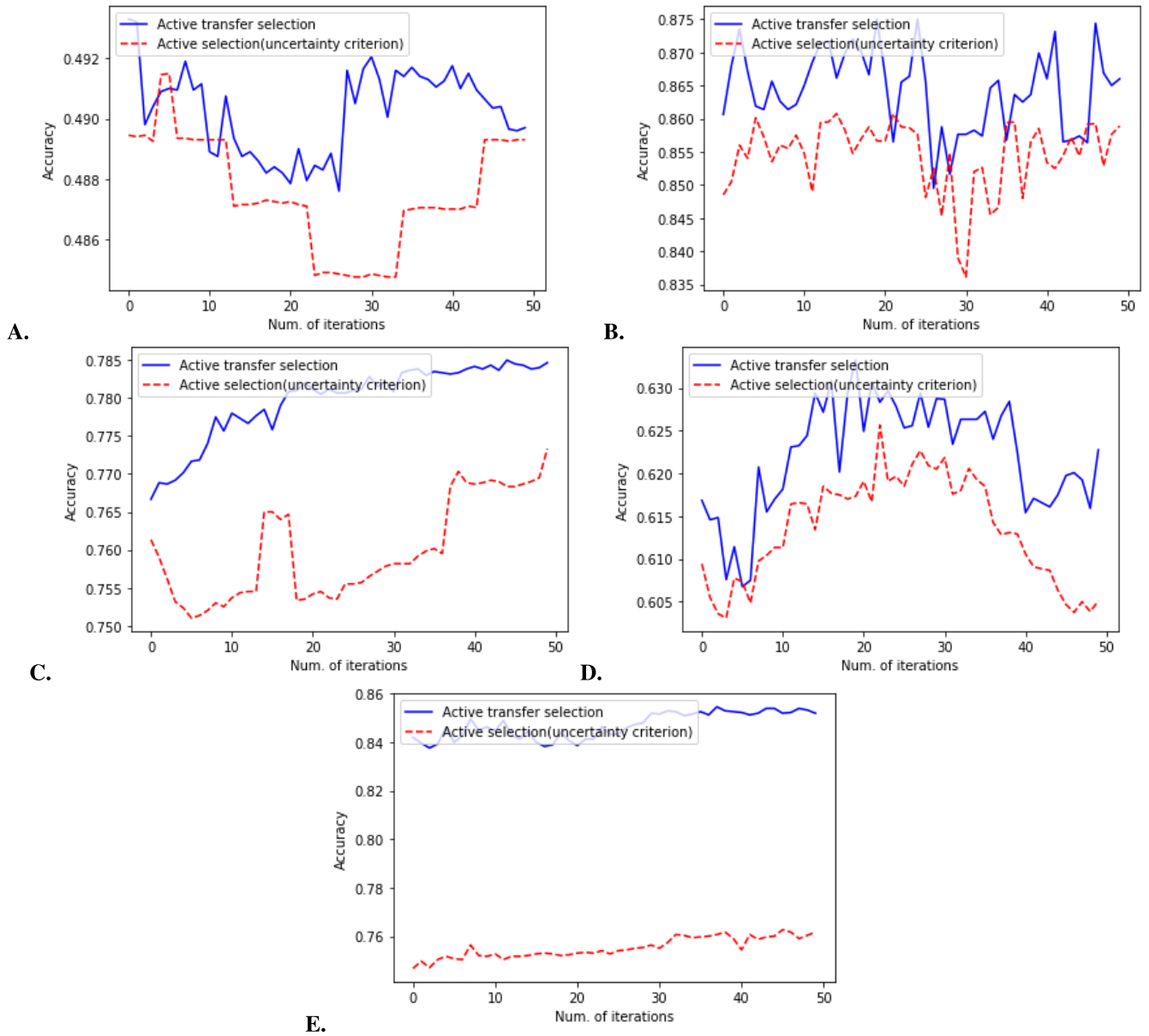
**FIGURE 4.** Comparison of precision obtained with random selection and with the two strategies of actively selecting training points: uncertainty sampling criterion and AT criterion. (A.) IMDB data set. (B.) Spam data set. (C.) Amazon product review data set. (D.) PadChest data set. (E.) Movie Review Polarity data set.

data set, the training set for the AT criterion contains only 4,500 records and the pre-trained data set was trained on only 1,574 records. The training set for the Movie Review Polarity data set consists of 1,500 records, and the pre-trained data set was trained on only 500 records.

To demonstrate the performance of the proposed method, the algorithm that performed the lowest was trained for each data set. The following algorithms were used for each data set: Decision Tree Classifier for the IMDB data set, GaussianNB for the Spam data set, Logistic Regression for the Amazon data set, Decision Tree for PadChest, and KNeighborsClassifier for the Movie Review Polarity data set. Figure 3 shows the comparison of the accuracy obtained

when the worst classifier was used for training and the points were randomly and actively selected using the uncertainty sampling criterion and the AT criterion for each data set used in the experimental evaluation.

Another way to demonstrate the efficiency of the proposed AT criterion is shown in Figure 4. The plots show the comparison of the precision metric obtained when the training points were randomly and actively selected using the uncertainty sampling and the AT criterion for each data set. The proposed AT criterion performs better for all data sets, but the efficiency of the method is more evident for the IMDB data set (Figure 4 A.) and for the Spam data set (Figure 4 B.).



**FIGURE 5.** Comparison of accuracy obtained using a deep learning model. The training points were selected using the two strategies of actively selecting training points: uncertainty sampling criterion and AT criterion. (A.) IMDB data set. (B.) Spam data set. (C.) Amazon product review data set. (D.) PadChest data set. (E.) Movie Review Polarity data set.

The plots from Figure 5 show the comparison of the accuracy obtained using a deep learning model to train the points which were selected using the two strategies of active selection: uncertainty sampling selection and AT criterion. As in the experiments presented before, we used five data sets for comparison. For each data set, when the strategy of selecting the training points is AT criterion, the algorithm performs better.

**IV. CONCLUSION**

By integrating active and transfer learning, this work investigated how to make an algorithm more efficient for classifying text data from small data sets. The proposed

criteria, which combines active and transfer learning, select those items that provide most of the information about the current task by using models learned on comparable tasks.

To confirm the potential of the proposed criteria, we considered three criteria for selecting the training points: random selection, uncertainty sampling selection and AT criteria. We found that there is a difference in performance for different data sets and different selection criteria, in particular, the proposed AT criterion performs better. The experimental results show that by combining active and transfer learning on a target domain, we can learn faster and with fewer labels than random selection.

In future work, we plan to further extend our approach by investigating the efficiency of the proposed method for multi-input models. Since in the paper [20] the proposed method was applied in image classification and in this work in text classification, the proposed method of selecting the training points by combining active and transfer learning can be integrated into models that use images and text as input.

## ACKNOWLEDGMENT

The author Daniela Onita is grateful to her baby who was very good inside and outside her womb and allowed her to work on her research projects. She loves her, baby R.

## REFERENCES

- [1] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering: New collection and results," in *Proc. 11th ACM Symp. Document Eng.*, Sep. 2011, pp. 259–262.
- [2] A. Birlutiu, P. Groot, and T. Heskes, "Efficiently learning the preferences of people," *Mach. Learn.*, vol. 90, no. 1, pp. 1–28, Jan. 2013.
- [3] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," *Med. Image Anal.*, vol. 66, Dec. 2020, Art. no. 101797.
- [4] H. Byun, J. Kim, D. Yoon, I.-S. Kang, and J.-J. Song, "A deep convolutional neural network for rock fracture image segmentation," *Earth Sci. Informat.*, vol. 14, no. 4, pp. 1937–1951, Dec. 2021.
- [5] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, "Active transfer learning network: A unified deep joint spectral–spatial feature learning model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1741–1754, Mar. 2019.
- [6] M. Enkhsaikhan, W. Liu, E.-J. Holden, and P. Duuring, "Auto-labelling entities in low-resource text: A geological case study," *Knowl. Inf. Syst.*, vol. 63, no. 3, pp. 695–715, Mar. 2021.
- [7] W. Fang, H. Luo, S. Xu, P. E. D. Love, Z. Lu, and C. Ye, "Automated text classification of near-misses from safety reports: An improved deep learning approach," *Adv. Eng. Informat.*, vol. 44, Apr. 2020, Art. no. 101060.
- [8] Y. Gautam, "Transfer learning for COVID-19 cases and deaths forecast using LSTM network," *ISA Trans.*, vol. 124, pp. 41–56, May 2022.
- [9] *Kaggle*. Accessed: Apr. 11, 2022. [Online]. Available: <https://www.kaggle.com/datasets/bittlingmayer/amazonreviews>
- [10] J. Kasai, K. Qian, S. Gurajada, Y. Li, and L. Popa, "Low-resource deep entity resolution with transfer and active learning," Jun. 2019, *arXiv:1906.08042*.
- [11] Y. Kim, Y. Kim, C. Yang, K. Park, G. X. Gu, and S. Ryu, "Deep learning framework for material design space exploration using active transfer learning and data augmentation," *NPJ Comput. Mater.*, vol. 7, no. 1, p. 140, Sep. 2021.
- [12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2011, pp. 142–150.
- [13] K. G. Ince, A. Koksai, A. Fazla, and A. A. Alatan, "Semi-automatic annotation for visual object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1233–1239.
- [14] B. Jiang, Q. Wu, X. Yin, D. Wu, H. Song, and D. He, "FLYOLOv3 deep learning for key parts of dairy cow body detection," *Comput. Electron. Agricult.*, vol. 166, Nov. 2019, Art. no. 104982.
- [15] B. Marinelli, M. Kang, M. Martini, J. R. Zech, J. Titano, S. Cho, A. B. Costa, and E. K. Oermann, "Combination of active transfer learning and natural language processing to improve liver volumetry using surrogate metrics with deep learning," *Radiol. Artif. Intell.*, vol. 1, no. 1, Jan. 2019, Art. no. e180019.
- [16] A. K. McCallum and K. Nigam, "Employing EM and pool-based active learning for text classification," in *Proc. Int. Conf. Mach. Learn. (ICML)*. Princeton, NJ, USA: Citeseer, Jul. 1998, pp. 359–367.
- [17] M. A. Al-Garadi et al., "Text classification models for the automatic detection of nonmedical prescription medication use from social media," *BMC Med. Inform. Decis. Making*, vol. 21, no. 1, pp. 1–13, 2021.
- [18] S. Moskal and S. J. Yang, "Translating intrusion alerts to cyberattack stages using pseudo-active transfer learning (PATRL)," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Oct. 2021, pp. 110–118.
- [19] M. R. Mohebbian, H. R. Marateb, and K. A. Wahid, "Semi-supervised active transfer learning for fetal ECG arrhythmia detection," *Comput. Methods Programs Biomed. Update*, vol. 3, 2023, Art. no. 100096.
- [20] D. Onita and A. Birlutiu, "Active learning based on transfer learning techniques for image classification," in *Proc. ESANN*, 2018, pp. 1–6.
- [21] D. Onita, A. Birlutiu, and L. P. Dinu, "Towards mapping images to text using deep-learning architectures," *Mathematics*, vol. 8, no. 9, p. 1606, Sep. 2020.
- [22] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.
- [23] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," Sep. 2004, *arXiv:cs/0409058*.
- [24] M. Rieker, A. Klein, F. Adrion, C. Hoffmann, and E. Gallmann, "Automatically detecting pig position and posture by 2D camera imaging and deep learning," *Comput. Electron. Agricult.*, vol. 174, Jul. 2020, Art. no. 105391.
- [25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.
- [26] B. Settles, "Active learning," *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 6, no. 1, pp. 1–14, Jun. 2012.
- [27] A. Singh and S. Chakraborty, "Deep active transfer learning for image recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–9.
- [28] H. Shi, H. Wang, C. Qin, L. Zhao, and C. Liu, "An incremental learning system for atrial fibrillation detection based on transfer learning and active learning," *Comput. Methods Programs Biomed.*, vol. 187, Apr. 2020, Art. no. 105219.
- [29] C. Schröder and A. Niekler, "A survey of active learning for text classification using deep neural networks," 2020, *arXiv:2008.07267*.
- [30] X. Tang, B. Du, J. Huang, Z. Wang, and L. Zhang, "On combining active and transfer learning for medical data classification," *IET Comput. Vis.*, vol. 13, no. 2, pp. 194–205, Mar. 2019.
- [31] C. Vununu, S.-H. Lee, and K.-R. Kwon, "A classification method for the cellular images based on active learning and cross-modal transfer learning," *Sensors*, vol. 21, no. 4, p. 1469, Feb. 2021.
- [32] X. Wang, T. K. Huang, and J. Schneider, "Active transfer learning under model shift," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2014, pp. 1305–1313.
- [33] D. Wu, B. Lance, and V. Lawhern, "Transfer learning and active transfer learning for reducing calibration data in single-trial classification of visually-evoked potentials," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2014, pp. 2801–2807.



**DANIELA ONITA** was born in Alba Iulia, Romania, in 1994. She received the Ph.D. degree in computer science from the University of Bucharest, Romania, in the long-distance studies program. She is currently an Associate Professor with the Department of Computer Science, "1 Decembrie 1918" University of Alba Iulia. Her research interests include machine learning, natural language processing, and computer vision.

• • •