Active Learning for Effectively Fine-Tuning Transfer Learning to Downstream Task

MD ABUL BASHAR and RICHI NAYAK, Queensland University of Technology

Language model (LM) has become a common method of transfer learning in Natural Language Processing (NLP) tasks when working with small labeled datasets. An LM is pretrained using an easily available large unlabelled text corpus and is fine-tuned with the labelled data to apply to the target (i.e., downstream) task. As an LM is designed to capture the linguistic aspects of semantics, it can be biased to linguistic features. We argue that exposing an LM model during fine-tuning to instances that capture diverse semantic aspects (e.g., topical, linguistic, semantic relations) present in the dataset will improve its performance on the underlying task. We propose a Mixed Aspect Sampling (MAS) framework to sample instances that capture different semantic aspects of the dataset and use the ensemble classifier to improve the classification performance. Experimental results show that MAS performs better than random sampling as well as the state-of-the-art active learning models to abuse detection tasks where it is hard to collect the labelled data for building an accurate classifier.

CCS Concepts: • Computing methodologies → Neural networks;

Additional Key Words and Phrases: Misogynistic tweet, hate speech, active learning, transfer learning, imbalanced dataset, topic model

ACM Reference format:
Md Abul Bashar and Richi Nayak. 2021. Active Learning for Effectively Fine-Tuning Transfer Learning to Downstream Task. ACM Trans. Intell. Syst. Technol. 12, 2, Article 24 (February 2021), 24 pages. 
https://doi.org/10.1145/3446343

INTRODUCTION

In the last decade, deep learning has become one of the most successful machine learning methods [41]. A deep learning model relies on the availability of a large dataset during training for learning the intricate features inherent in the dataset; hence, using a small dataset in training the model usually results in an overfitted model [41, 62]. Transfer learning (TL) is commonly used to deal with the problem of data scarcity. Transfer learning is a process of acquiring knowledge in one problem domain and applying it to a different but related problem [50, 62, 80]. For instance, the knowledge gained while learning to classify cats and dogs can be useful when classifying tigers and lions.

Recently, Language Model (LM) has become a popular method of transfer learning in Natural Language Processing (NLP) tasks [29, 41, 62]. An LM is pretrained using a large general-purpose...
unlabelled text corpus and is fine-tuned with the labelled data available for a downstream target task. Several applications such as text classification [41], next sentence prediction [29], question answering, and natural language inference [62] have been developed using an LM in the TL fashion.

An LM is designed to capture the linguistic aspects of a corpus during its pretraining. It can be biased to the features present in that corpus and may adversely affect transfer learning in another problem [41, 50]. We conjecture that exposing the LM model to instances selected based on multiple aspects of the dataset in the fine-tuning step will better generalise the LM model. Example aspects that can be learned are topics, language, semantic relations, hierarchical structures, ontological concepts, and many more.

In this article, we propose a framework that enables the inclusion of multiple diverse aspects in the instance selection process to improve classifiers’ accuracy. We investigate how to sample a combination of topic and language-based instances to fine-tune a pretrained language model for improved generalisation when applying to a downstream classification task. We propose to sample instances in an active learning setting.

Active learning is a popular method of selecting informative instances in a data distribution to effectively build a classifier [49, 64, 68, 69]. Active learning techniques vary in how they sample informative instances, for example, by selecting the instances near a discriminating boundary [17, 46, 65] or instances for which classifiers in an ensemble disagree the most [67]. It is assumed that the classifier has less knowledge about those instances, and including them in training will update its knowledge. These techniques are known to suffer from sampling bias because the instances may have a different distribution than the original dataset [20, 26, 78].

Another family of active learning techniques addresses this problem by utilising feature distribution structures in instance selection [20, 26, 78]. For example, representative instances have been selected from each cluster present in the data [26, 78]. The performance of this approach depends on how well the clusters have been identified in the dataset [42]. In a document collection, the instances are assigned into clusters based on shared features (i.e., common words shared amongst documents). The underlying assumption is that each cluster will be pure (i.e., members of a cluster will belong to the same class). This approach may not be effective in applications where two classes share highly similar features. For example, in the misogynistic abusive tweet detection task, misogynistic abusive keywords appear in both misogynistic and non-misogynistic classes [11, 12]. Clustering alone may not be able to effectively separate instances that share the same text but differ in meaning/context.

To select informative and diverse instances, we need to consider specific semantics aspects (e.g., topics, linguistic features, hierarchical structure, semantic relations or ontology) in the dataset. None of the existing techniques concentrate on selecting instances by focusing on specific semantic aspects of a dataset. More importantly, in addition to focusing on the semantic aspects, it needs to control the number of instances sampled from each class; otherwise, instances of the minority class may hide in the crowd of majority class instances.

We propose a Mixed Aspect Sampling (MAS) framework that enables sampling of instances representing different aspects of the dataset and builds an ensemble of classifiers where each classifier can focus on a specific semantic aspect in the dataset. We implement the MAS framework with two aspects: topical and linguistic. One classifier is trained to label the text data from the perspective of topical aspects, and another classifier is trained to label the data from the perspective of linguistic aspects.

We propose to deploy MAS in batch-mode fashion [64, 69, 84] to deal with the downstream task of the scarce labelled dataset. In early batches, classifiers in the ensemble model disagree more about the label of the instances as they lack knowledge of instances that can be differently labelled
when seen from other perspectives. We use a cluster-based down-sampling to select instances from three groups: (a) instances where classifiers disagree about the labels, (b) instances where classifiers agree that the labels belong to the majority class, and (c) instances where classifiers agree that the labels belong to the minority class. As the number of batch passes, the agreement between classifiers increases as they gain knowledge from the sampled instances where they disagreed. By controlling the down-sampling for each group, we can control the recall for a minority class (i.e., more down-sample the majority class and less down-sample the minority class).

We apply the MAS framework in two downstream tasks, misogynistic tweet detection and hate tweet detection, where obtaining the labelled tweets is a rigorous task. Abuse detection on Twitter has attracted considerable attention in recent years. An automated misogynistic/hate abuse detection system could improve the understanding of the patterns, driving factors, and effectiveness of responses over a sustained period [11, 12]. However, for such detection, training a neural network-based classifier with a small set of labelled data is difficult due to the complex nature of tweets and the vast number of parameters in neural network models [11, 12].

We conducted experiments on datasets collected from Twitter to evaluate the performance of MAS. Experimental results show that exposing the LM model to instances selected by MAS during the fine-tuning step generalises the LM model better than exposing the LM model to instances selected by state-of-the-art active learning models such as Marginal Probability Distribution Matching (MPDM) [20], HSAL [26], USM [65], QBC [67], and LLAL [83]. In other words, MAS selected diverse instances can improve the accuracy of the LM model more than the state-of-the-art active learning models.

The specific contribution of this article is three-fold. (a) It proposes a MAS framework to sample instances representing different semantic aspects (e.g., topic and language) of the dataset for active learning. (b) It proposes a transfer learning model using these diverse sampling instances to two downstream target tasks of abuse detection. (c) It gives a theoretical analysis of topic-focused and language-focused classifiers using the random set [37, 45, 56] and Bayesian probability distribution.

The rest of the article is organised as follows: Section 2 discusses related work, Sections 3 and 4 present MAS and its empirical evaluation, respectively, and Section 5 concludes the article.

2 RELATED WORK

This section covers related works in the areas of active learning and transfer learning with language models. It also includes a brief discussion on abusive tweet detection that is used as a downstream classification task, to show the effectiveness of the MAS framework of active learning on top of the language model in a transfer learning setting.

2.1 Active Learning

Active learning is commonly used in situations where the unlabeled instances are abundant but their manual labelling is expensive. An active learning algorithm interactively chooses training instances and queries an oracle (i.e., a human) to label them with the desired outputs [20, 26, 65, 67]. The algorithm chooses the training instances such that the number of required instances for learning a concept is much lower than the number required in the normal supervised learning setting.

Active learning is widely used in many areas such as text classification [75], information extraction [74], image classification [87], and speech recognition [85]. There exist several active learning approaches. Approaches based on uncertainty sampling [17, 46] and ensemble modelling [67] use classifiers to select informative instances. However, these approaches often suffer from the sampling bias because the selected instances may have a different distribution from the original data.

To address this problem, in a recent approach [26], a density clustering-based method [78] is used to select representative instances appearing in each cluster. The performance of this
approach depends on the quality of clustering results [42]. Additionally, when a cluster-based model is used alone, it may fail in datasets where instances of different classes share the same features [11, 12]. Clustering-based sampling may perform poorly when instances do not exhibit a clear cluster structure [18]. For example, tweets expressed using same/similar text may mean something different in different contexts. People commonly use abusive text in expressing normal tweets, especially within their close circle of friends or expressing content in a satirical or jovial manner. This approach may not be effective in applications where two classes share highly similar linguistics features.

A possible solution to overcome this problem is the consideration of diverse semantics aspects for choosing the instances. There exist a handful of active learning models [49, 64] that emphasise on selecting diverse instances. Their goal is to avoid selecting similar instances for annotation that might negatively affect the performance. For example, a similarity matrix constructed using original features was utilised to find diverse instances [49]. However, the features used represented the same aspect of instances. Similarly, authors in [64] used Jeffrey divergence (to measure the distance between two probability distributions) in feature space of the same aspect for selecting diverse instances. To the best of our knowledge, none of the existing works considers diverse (i.e., multiple semantic) aspects in sampling instances; they may use diversity in the same aspect. The MAS framework proposes a new approach of active learning by introducing diverse aspects (e.g., linguistic, topical, semantic relations.) in the instance selection process to improve the accuracy of classifiers.

Another issue faced by active learning methods such as random sampling and uncertain sampling is failing to select minority class examples in imbalanced data distribution [5]. Guided learning, based on crowd-sourcing to find or generate class-specific training instances, can help to get more balanced class frequencies [4, 58]. However, guided learning is resource consuming and may not present the true distribution generating training examples. Sometimes, heuristic labelling methods such as distant supervision [24] or data programming [31] are used for datasets with the imbalanced class distribution. However, these methods are only applicable when a good knowledge base or pretrained predictor is available [16].

To address this problem, the proposed MAS framework categorises instances into three groups and applies cluster-based down-sampling in each group to make a balance between the instances selected for classes.

2.2 Abusive Tweet Detection

Lack of unique and discriminative linguistic characteristics in abusive tweets such as misogynistic and hate speech makes their separation from non-misogynistic and non-hate speech difficult [79, 89]. Offensive words or expletives are commonly used on social media to express satire and jokes [79]. Machine learning models need to incorporate the context of misogynistic and abusive words in learning instead of just relying on the occurrences of those words. For example, the occurrence of a misogynistic word such as slut in the tweet life is a slut does not make the tweet misogynistic.

Some works used syntactic features and the intensity of hate speech in classification [1, 36, 70]. In AMI@IberEval shared task-A [1], to detect misogynistic tweets, participants used individual algorithms as well as ensembles based on Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and Stochastic Gradient Descent. Researchers explored traditional techniques (i.e., bag-of-words/characters (BOW) and unigrams and bigrams) as well as specific lexical features for representing the tweet contents. Methods that spent efforts on manual feature engineering were able to achieve improved performance. Emphasis on feature representation was found more significant than the use of an advanced machine learning method.
The highest performing team used a vector representation that concatenates sentence embedding, weighted BOW, and average word embeddings coupled with a Logistic Regression model. Performance of traditional algorithms highly depends on feature engineering and feature representation [27, 82]. NN-based classifiers have become popular due to their ability to automatically learn abstract features from the given input feature representation and reduced dependency on manual feature engineering [6]. For example, a Deep Neural Network (DNN) model has shown to extract discriminative features that can capture the semantics of hate speech [89]. Popular NN models used in this area are Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM). In AMI@IberEval shared task-A [1], a DNN was adopted by a team that achieved a moderate performance. An ensemble of RNN classifiers has been used to identify hateful content in social media [59].

NN classifier models typically require a huge amount of labelled data for training, otherwise they can overfit the training data [12, 73]. Random dropout [72], L1 and L2 regularisation are commonly used to regularise NN architecture. These regularisations reduce overfitting for medium size datasets. NN models overfit to small datasets [12, 34]. Researchers have used transfer learning to address this problem. For example, to identify misogynistic abuse, [11] proposed a progressive transfer learning through a number of unlabelled datasets to reduce overfitting of a classification model trained with a small labelled dataset. However, it still must have a quality dataset available for fine-tuning to the downstream classification task. The MAS framework enables generation of datasets available for fine-tuning these classification tasks.

2.3 Transfer Learning

The simplest form of transfer learning in text processing has been used by generating features from the pretrained word embeddings (also called pretrained word vectors) [76], sentence embeddings [51], and paragraph embeddings to be used in the downstream model of an NLP task. These pretrained embeddings provide an improvement over embeddings learned from only the target data [12].

An emerging trend is pretraining a language model (LM) on a related large scale dataset [12, 25, 29, 41, 62]) and then fine-tune the model to a supervised downstream task such as classification. The top-performing team named NULI [48] in SemEval-2019 [48] used the pretrained transformer (BERT) for detecting offensive language on social media.

The pretraining of an LM involves learning the linguistic aspects of the dataset used in training. This process may make the LM model be biased to the linguistic aspects appearing in the dataset used in training. Applying this model to a downstream supervised task in a straightforward fashion to fine-tune may result in adverse transfer learning due to mismatch of aspects. We propose the MAS framework that exposes the LM model to instances, selected based on multiple aspects of the dataset, for fine-tuning to generalise the model better.

MAS proposes how to sample a mixture of instances that capture different semantic aspects of a dataset to fine-tune a pretrained LM. There might be several aspects to be considered; for example, topics, language, hierarchical structures, ontological concepts, and so forth. This article implements a model to sample a mixture of topic and language-focused instances.

3 THE PROPOSED MAS FRAMEWORK

The overall aim of MAS is to provide the best labelled training set for fine-tuning an LM-based model to a downstream target task (e.g., classification) where it is difficult to get a labelled training dataset. It is expensive and time-consuming to request the oracle (i.e., human labeller) for obtaining a class label for each instance. MAS obtains a quality training set covering diverse aspects inherent
in the dataset but by reducing the number of requests made to the oracle for labeling only the selected instances.

A transfer learning model based on LM is pretrained to learn linguistic aspects from a large general-purpose text collection. As a result, such transfer learning might not capture other semantic aspects or be biased to the linguistic aspect. Most of the existing active learning techniques select the instances where the classifier is most uncertain of the class labels [17, 46, 67]. However, when selecting instances, these techniques do not focus on any specific semantic aspects such as topics, language, hierarchical structures, ontological concepts, and so forth. As result, fine-tuning an LM-based transfer model for a downstream classification task with the instances selected by existing active learning techniques cannot address the semantic bias problem. MAS proposes to address this problem by focusing on specific semantic aspects when selecting the instances.

### 3.1 Overview

Figure 1 illustrates the MAS framework that follows an active learning setting [17, 46, 67, 68]. MAS is deployed in the batch-mode fashion [84] that aims to obtain a diverse set of instances that represent different aspects in the collection. MAS combines two classification models where one model labels tweets focusing on topical aspects and the other model labels tweets focusing on linguistic aspects.

The ensemble model categorises the instances into three groups: (1) agreed majority class by both models, (2) agreed minority class by both models, and (3) disagreed (or uncertain) instances where the models do not agree on the labels of the instances. The instances from each group are down-sampled using a clustering technique (e.g., Hierarchical Agglomerative Clustering).

For each batch of instances that are categorised into one of three groups by the ensemble model and then selected by the cluster-based down-sampling, MAS consults the oracle to obtain the class labels of those sampled instances. These sampled instances along with oracle provided labels are added to the training set to update the models. This process is repeated until the framework gathers a reasonably sized dataset or a dataset that can provide the required level of performance. This labelled instance set, reflecting the diverse aspect in the collection, becomes an
input to fine-tune the LM-based transfer learning model for the downstream classification task (e.g., abusive tweet detection).

### 3.2 Aspect-Focused Sampling

As the tweets selected after each batch are checked by an oracle, only the most informative tweets of each batch should be passed on to oracle to improve the feasibility and efficiency of the process. The MAS framework enables this by using the ensemble of classifiers and cluster-based down-sampling for choosing selective instances. For example, instances where the classifiers in an ensemble disagree about the labels are more likely to be informative [66]. Using the cluster-based down-sampling to select some of these instances makes them less likely outliers [90]. In this section, we discuss how to induce specific semantic aspects from underlying data distribution to the classifier through aspect-focused disagreement of classes and down-sampling by clustering.

The MAS framework (as shown in Figure 1) uses an ensemble of two classifiers. One classifier based on a topic model (i.e., topic-focused classifier) emphasises on the topical features in the dataset during sampling tweets for each class. It uses the topical knowledge inherent in the dataset to label tweets in classes. A detailed description of this classifier is given in Section 3.3. Another classifier based on a language model (i.e., LM-focused classifier) emphasises on using the linguistic attributes (such as semantics from word co-occurrence and structure from word sequence and order) during sampling tweets for each class. It uses linguistic knowledge to label tweets in classes. A detailed description of the language-focused classifier is given in Section 3.4.

For each tweet in a batch, the ensemble model is applied to check if the two classifiers agree on the tweet class. All the tweets in a batch are grouped into three categories, namely, **agreed-minority-class**, **agreed-majority-class**, and **disagreed or uncertain**.

Each classifier in the ensemble model is trained to focus on a specific aspect and there are some instances for which models will disagree when they present a different perspective. In other words, models will disagree when they do not have enough knowledge about these instances. Table 1 shows three instances in three categories. Models need to learn this knowledge gap. We propose to sample most instances from this group so that the models can have more knowledge about the conflicting instances. For example, a classifier focusing on the linguistic aspect will have more instances to learn where it disagrees with the classifier focusing on the topical aspect. The assumption is that the knowledge a classifier gets from these instances will make them more probable to agree in the future.

We sample a smaller portion from the instances where the models agree on the labels based on the assumption that the model has knowledge of those characteristics. We continue to sample from the group of correctly labelled instances to reinforce the model learning. We separate these instances into two groups (i.e., agreed majority and minority classes) to control sampling in imbalanced datasets where one class of instances dominates others in number. This separation allows MAS to collect sufficient instances for the minority class.
We propose to use a cluster-based down-sampling to select the required instances from each group. We use an agglomerative hierarchical clustering [28] to find the clusters and randomly selected instances from each cluster. By controlling the number of clusters in each group, we can control the recall for the corresponding class. For a given group, reducing the number of clusters means more generalisation of knowledge to aggregate the instances within the smaller number of clusters. On the other hand, increasing the number of clusters means allowing more details of the knowledge that can accommodate the larger number of clusters. As an instance is randomly (or some other way) selected from each cluster, an increasing number of clusters also means selecting more instances for a group. In other words, we can feed more detailed knowledge and training instances for a group by increasing the number of clusters, which will allow the classifier to know more of the corresponding class and increase its recall. Recall controlling can significantly contribute to improving a classifier with underlying imbalanced data distribution or minority classes.

The selected instances are manually checked and corrected for class assignment by an oracle (i.e., human expert). Finally, the instances with their corrected class labels are added to the labelled tweet set. The labelled tweet set is used to train the LM-focused and the topic-focused classifiers. The MAS framework iterates in a batch mode active learning fashion, i.e., the framework starts with a very small labelled dataset that it uses to train the classifiers. The classifiers are then used to sample a batch of instances for manual labelling. The manually labelled instances are added to the labelled dataset to increase its size. The iteration continues until we get a satisfactory labelled dataset or the labelling budget is finished.

For aspect-focused classification, we need to incorporate a mechanism in the classifier training that will allow the classifier to focus on a specific aspect. In the remainder of this section, we formalise how to incorporate a specific aspect in the classifier training. Let $t = (f_1, \ldots, f_n)$ be a vector representing a data instance with $n$ features. Let $C_k$ be a set of $K$ classes. The classification task is to assign an instance to a class $C_k$ based on the feature vector $t$, i.e., finding $p(C_k | t)$. A generative model (e.g., Naive Bayes) learns the probability distribution $p(t | C_k)$ and $p(C_k)$ that can be transformed into $p(C_k | t)$ by Bayes theorem [30]:

$$p(C_k | t) = \frac{p(t | C_k) p(C_k)}{p(t)} = \frac{p(t, C_k)}{p(t)}.$$  

(1)

Probability $p(C_k)$ in Equation (1) is prior knowledge, and generative models sometimes use it to regularize the model [15]. Equation (1) can be rearranged as

$$p(t, C_k) = p(C_k | t)p(t).$$  

(2)

Probability $p(t)$ in Equation (2) can be seen as a regulariser for $p(C_k | t)$. That is, $p(t)$ can regularise joint probability distribution $p(t, C_k)$. The distribution $p(t)$ informs the probability of observing $t$ in a population, while $p(C_k | t)$ informs the probability of the class $C_k$ being observed when $t$ is observed. In other words, the probability of a randomly chosen instance $t$ being in a population (e.g., in a collection of tweets, sequences, linguistic structures, topics, or concepts) is $p(t)$. The probability of $t$ belonging to class $C_k$ is $p(t, C_k)$. That is, $p(t, C_k)$ can capture both the classification and the population distribution of a specific aspect (i.e., topics, linguistic structure, etc.).

A discriminative model (e.g., SVM, LSTM) learns to classify a data instance $t$ into class $C_k$ (i.e., it learns the conditional probability distribution as $p(C_k | t, \theta) \approx p(C_k | t)$, where $\theta$ is the list of model parameters). In a discriminative model, the least certain instance lies near the classification boundary [20, 46, 63]. Many existing active learning techniques use this least certain instance as an informative sample for classifier [46, 63]. However, this instance is not necessarily a representative of different aspects in the dataset, so knowing its label is less likely to improve accuracy on the
data as a whole [65], especially when a pretrained transfer learning model is biased to a specific aspect (e.g., linguistic aspect).

It is more useful to learn probability $p(t)$ of an instance $t$ from the different aspects (e.g., language, topics) in the underlying data distribution and plug it into Equation (2) so that the classifier can focus on various aspects. For example, we can learn $p(t)$ from a language model and plug into Equation (2) to make a classifier focus on linguistic aspects and learn $p(t)$ from a topic model and plug into 2 to make a classifier that focuses on topical aspects. It is to be noted that $p(C_k | t)$ is associated with the class label, whereas $p(t)$ does not. This ascertains that $p(t)$ can be learned independent of the class label (i.e., in unsupervised training). In the following section, we discuss the topic-focused classifier, which is followed by the section discussing the language-focused classifier.

### 3.3 Topic-Focused Classification (TFC) Model

The goal of using topic-focused sampling is to emphasise the topics in the collection when labelling it for a classification problem. By focusing on the topics in the collection, we have a higher probability of capturing the underlying topics that are discussed in the instances of a given class. The topic-focused classification (TFC) model will learn what are the topics in data distribution that are discussed predominately in each class. For example, the collection on global warming can have topics like pollution, greenhouse gas, ozone layer depletion, and so forth. Examples of topics in misogynistic tweets are harassment, threat, violence, promotion of self-harm, suicide, hateful conduct, making sexual advance, objectifying a person, and so forth. Our previous study [2, 35] shows that the topic-focused information filtering can find relevant user information needs.

The topic-focused classification process is shown in Figure 2. Topics are discovered from the text collection using a topic modelling technique and the topical weight of each feature in the dataset is estimated using topics through Random Set [37, 45, 56]. As topic modelling (e.g., LDA [14]) cannot capture the class polarity of features, we use a discriminative classifier (e.g., SVM) to estimate the polarity weight of each feature. Discriminative classifier finds a class boundary for the dataset [39]. We estimate the polarity weight of a feature by measuring its distance from the class boundary. We combine topical and polarity weights by scaling them together to estimate feature weights. Finally, the weighted feature vectors are used to classify instances (e.g., summing the feature weight of an instance [7–10, 35]). We use weighted feature vectors to learn probability $p(t)$ in Equation (2) from the topical perspective.

We use Latent Dirichlet Allocation (LDA) [14] to discover hidden topics in a collection due to its popularity for finding latent topics [3, 22, 43]. However, any model adapted to discover topics in a dataset will suffice here. LDA is defined as a multinomial distribution over features in a corpus [14]. The majority of text mining models including LDA use the bag-of-words model for feature representation (i.e., features are conditionally independent of each other). Therefore, to
estimate feature weights, we can simplify Equation (2) for this representation using Naive Bayes Assumption.

3.3.1 Naive Bayes Assumption. Given a class $C_k$, Naive Bayes assumes that each feature $f_i$ is conditionally independent of every other feature $f_j$ for $i \neq j$. Based on this assumption, Equation (2) can be rewritten as

$$p(t, C_k) = p(C_k | t)p(t) = p(C_k | f_1, \ldots, f_n)p(f_1, \ldots, f_n) = p(C_k | f_1, \ldots, f_n)p(f_1 | f_2, \ldots, f_n) \ldots p(f_{n-1} | f_n)p(f_n) \approx p(C_k | f_1, \ldots, f_n)p(f_1)p(f_2) \ldots p(f_n) \approx p(C_k | f_1)p(C_k | f_2)p(f_1)p(f_2) \ldots p(f_n) = p(C_k | f_1)p(C_k | f_2)p(f_2) \ldots p(C_k | f_n)p(f_n) = \prod_{i=1}^{n} p(C_k | f_i)p(f_i) \propto \sum_{i=1}^{n} \ln[p(C_k | f_i)p(f_i)] = \sum_{f \in t} \ln[p(C_k | f)p(f)] \propto \sum_{f \in t} p(C_k | f)p(f).$$

For a given feature $f \in t$, $p(f)$ is independent of $C_k$. This means $p(f)$ can be interpreted as the probability of observing the feature $f$ in the population, while $p(C_k | f)$ is the probability of the class $C_k$ being observed when $f$ is observed. We model polarity weight of a feature $f$ as $p(C_k | f)$ estimated using a discriminative classifier SVM; and we model topical weight as $p(f)$ estimated through Random Set and the topics discovered by LDA. We discuss SVM and show how to estimate polarity weight $p(C_k | f)$ in the following section.

3.3.2 SVM. We use a linear SVM [39] for learning the class polarity of features because SVM-based active learning has been found effective in prior research [32]. As shown in Figure 3, SVM is a discriminative classifier that uses support vectors to find the hyperplane $a$ that best separates two classes. The normal distance to the hyperplane (i.e., the distance of a feature from the hyperplane) is used to estimate the class polarity $p(C_k | f)$ of a feature $f$. In the following section, we discuss LDA to learn hidden topics of a tweet batch.

3.3.3 LDA. Let $T = \{t_1, t_2, \ldots, t_M\}$ be a batch of $M$ instances. Each instance $t$ is a bag of features (e.g., words). Let $F = \{f_1, f_2, \ldots, f_V\}$ be the set of unique features in the tweet batch $T$, where $V$ is the size of the vocabulary in that batch. The idea behind LDA is that features observed in each tweet are generated by a tweet-specific mixture of batch-wide hidden topics [14]. The number of hidden topics is assumed to be fixed to $N$.

A topic $z_j$ is represented as a multinomial probability distribution over the $V$ features as $p(f_j | z_j)$, where $1 \leq j \leq N$ and $\sum_{j=1}^{V} p(f_j | z_j) = 1$. A tweet $t$ is represented as a probabilistic mixture of topics as $p(z_j | t)$. Therefore, the probability distribution of the $i$th feature in tweet $t$ can be modelled as a
Fig. 3. Estimating polarity weights of features using SVM, where \( a \) is maximum-margin hyperplane, \( b_1 \) and \( b_2 \) are margins of the SVM, and \( x_1 \) and \( x_2 \) are feature dimensions.

Fig. 4. Set-valued mappings of features (\( F \)), tweets (\( T \)), and topics (\( Z \)).

mixer over topics:

\[
p(f_i|t) = \sum_{j=1}^{N} p(f_i|z_j)p(z_j|t).\]

Here, the only observable variable is \( p(f_i|t) \). The other two variables, \( p(f_i|z_j) \) and \( p(z_j|t) \), are hidden. We use the widely used statistical estimation technique \([19]\) to learn these hidden variables.

The probability distribution \( p(f) \) estimated from underlying topic distribution is our topical weight of the feature \( f \). A batch consists of multiple text (tweet) instances. Each text instance has multiple features and can discuss multiple topics. However, we do not know how each topic is relevant to others. We model the complex relationships between features, tweets, and topics using the Random Set theory \([37, 45, 56]\) and use this modelling to estimate \( p(f) \).

3.3.4 Topical Weights of Features. Let each tweet \( t \in T \) be a probabilistic distribution over the feature space \( F \), which can be modelled using the set-valued mapping \( \Gamma_1(t) \). We assume that representativeness of a feature \( f \) in the random tweet space is a probabilistic mixture of \( T \), which can be modelled using the inverse set-valued mapping \( \Gamma_1^{-1}(f) \). The mappings are graphically shown in Figure 4.

The set \( T \) can be considered as the evidence space because a set of features represents a tweet \( t \). However, a feature’s representativeness level in the evidence space is unknown. The probability distribution \( \Psi_1 \) is defined on \( T \) to indicate this uncertainty.
Let the probability of feature \( f \) be the representative of \( t \) as \( p(f|t) \), where, for simplicity, we assume \( p(f|t) = 1 \) if \( f \in t \) and \( p(f|t) = 0 \) if \( f \notin t \).

We use the random set theory to model \( \Psi_1 \). A random set is an arbitrary entity that contains a subset of values selected from a given space [57]. If there is a probability distribution \( \Psi_1 \) defined on the evidence space \( T \), then the pair \( (\Psi_1, \Gamma_1) \) is called a random set [37, 45, 56].

With each tweet \( t \) described by the probability distribution over the feature set \( F \), the relationship between a set of features and a tweet is represented and described by the set-valued mapping as

\[
\Gamma_1 : T \rightarrow 2^F - \{\emptyset\}; \Gamma_1(t) = \{ f \in F | p(f|t) > \zeta \},
\]

where \( \Gamma_1(t) \subseteq F \) for all \( t \in T \) and \( \zeta \) is a user-defined threshold, set to \( \zeta = 0 \) for simplicity in this research.

As there is a need to identify the relevance level of feature \( f \), the inverse set-valued mapping of \( \Gamma_1 \) is considered to estimate a representative distribution \( \Psi_1 \) on \( T \). For all features \( f \in F \), the relationship between a feature and a set of tweets is represented and described by the inverse set-valued mapping of \( \Gamma_1 \) as

\[
\Gamma_1^{-1} : F \rightarrow 2^T; \Gamma_1^{-1}(f) = \{ t \in T | f \in \Gamma_1(t) \}.
\]

The pair \( (\Psi_1, \Gamma_1^{-1}) \) models an underlying data distribution in the tweet space for features. Therefore, the representativeness weight \( w_t(f) \) of feature \( f \) to the tweet space can be estimated as follows:

\[
p(f) = \Psi_1(f) \propto w_t(f) = \sum_{t \in \Gamma_1^{-1}(f)} p(f|t) \times p(t), \tag{4}
\]

where \( p(t) \) is the probability of \( t \) being representative of the underlying data distribution in the tweet space.

As tweet \( t \) can contain multiple topics, we assume that \( t \) is a probabilistic mixture of a set of topics \( Z \) in \( T \), which can be modelled using the set-valued mapping \( \Gamma_2(t) \). The mappings are graphically shown in Figure 4. The set \( Z \) is the evidence space in this case and can infer the representativeness of \( t \) by the topics that users discuss in the tweet space. However, the representative level of \( t \) is unknown. Similarly, as before, \( \Psi_2 \) is a probability distribution defined on \( Z \) to indicate this uncertainty. The pair \( (\Psi_2, \Gamma_2) \) is called a random set in this case.

Let the probability of tweet \( t \) being the representative of a given topic \( z \) be \( p(t|z) \). As each tweet \( t \) is described by the probability distribution over the set of topics, a set-valued mapping of \( \Gamma_2 \) is defined as

\[
\Gamma_2 : T \rightarrow 2^Z - \{\emptyset\}; \Gamma_2(t) = \{ z \in Z | p(t|z) > \xi \},
\]

where \( \Gamma_2(t) \subseteq Z \) for all \( t \in T \) and \( \xi \) is another user-defined threshold assigned to \( \xi = 0.01 \), as the default value of Gensim LDA.\(^1\) Thus, the representativeness of \( t \) of topics in tweet space can be estimated as follows:

\[
p(t) \propto \Psi_2(t) \propto \sum_{z \in \Gamma_2(t)} p(t|z). \tag{5}
\]

Using Equation (4) and Equation (5), the representativeness weight \( w_t(f) \) of the feature \( f \) can be calculated as follows:

\[
p(f) \propto w_t(f) = \sum_{t \in \Gamma_1^{-1}(f)} \left( p(f|t) \times \sum_{z \in \Gamma_2(t)} p(t|z) \right). \tag{6}
\]

As Figure 4 shows, this model finds tweets that are representative of topics, then finds features that are representative of those tweets and topics.

\(^1\)https://radimrehurek.com/gensim/models/ldamodel.html.
To find the latent topics in $T$, we use LDA that gives us $\rho(z|t)$, but we need $\rho(t|z)$. Therefore, by applying Bayes’ theorem, we can get $\rho(t|z) = \frac{\rho(z|t) \times \rho(t)}{\rho(z)}$. Here, $\rho(t)$ is a prior distribution that can be ignored, and $\rho(z)$ is the marginal probability of $z$ in $T$.

3.4 Language-Focused Classification Model (LM)

The goal of using language-focused sampling is to emphasise the linguistic aspect of the dataset (e.g., semantics from word co-occurrence and structure from word sequence and order) when sampling tweets for each class. By focusing on the linguistic attributes of tweets, we have a higher probability of capturing the underlying language tones and structures that are prevalent in a given class. In this section, we formalise how to incorporate linguistic aspect in the classifier training.

A LM is represented as a probability distribution over sequences of words. For example, given a sequence of $n$ words $t = (f_1, \ldots, f_n)$, it assigns a probability $p(f_1, \ldots, f_n)$ to the whole sequence. We show next that, using the chain rule, the joint probability in the left-hand side of Equation (2) can be rewritten as

$$p(C_k | t)p(t)
= p(C_k, t)
= p(C_k, f_1, \ldots, f_n)
= p(C_k | f_1, \ldots, f_n)p(f_1, \ldots, f_n)
= p(C_k | f_1, \ldots, f_n)p(f_1 | f_2, \ldots, f_n)p(f_2, \ldots, f_n)
= \ldots
= p(C_k | f_1, \ldots, f_n)p(f_1 | f_2, \ldots, f_n) \ldots p(f_{n-1} | f_n)p(f_n)
= p(C_k | f_1, \ldots, f_n) \prod_{i=1}^{n} p(f_i | f_1 \ldots f_{i-1})$$

(7)

The factor $\prod_{i=1}^{n} p(f_i | f_1 \ldots f_{i-1})$ in Equation (7) is similar to a language model. This means we can learn $p(t) = \prod_{i=1}^{n} p(f_i | f_1 \ldots f_{i-1})$ using a language model and plug into Equation (2).

Utilising word co-occurrence and sequence, an LM can encode the complexities of a language such as semantics and grammatical structure, as well as, it distils a fair amount of knowledge from a corpus [44]. In other words, when an LM is trained on a vast amount of data, it can learn useful knowledge encoded in the training data [88]. For example, it can capture long-term dependencies [47], hierarchical relations [38], and sentiments [61].

We propose to use a classifier trained with a pretrained Neural Network Language Model (NNLM) to sample a given tweet batch focusing linguistic aspect. Next, we show how an NNLM is used in the MAS framework.

3.4.1 NNLM. A language model designed with a neural network to provide a probability distribution over sequences of words is called NNLM [44]. Given a sequence of $n$ words $t = (f_1, \ldots, f_n)$ as input, it provides a probability $p(f_1, \ldots, f_n)$ as output. RNNs and its variants have been commonly used as NNLM [21, 44, 54, 55, 81, 86]. LSTM [40], an advanced variant of RNN, uses a gating mechanism to ensure proper propagation of information through many steps of a sequence. It has provided state-of-the-art performance for several benchmark language modelling tasks [44, 52, 53].
Figure 5 shows a NNLM architecture based on LSTM. Similar to a LM, RNN- or LSTM-based NNLMs employ the chain rule to model joint probabilities over the word sequences:

\[ p(t) = p(f_1, \ldots, f_n) = \prod_{i=1}^{n} p(f_i | f_1 \ldots f_{i-1}). \]  

Here, for example, the context of all previous words in the sequence is encoded within an LSTM, and the probability of getting next word is distributed over the words (in vocabulary) using a Softmax function [44].

RNN-based models and its variants use a text sequence either from left to right or concatenated left-to-right and right-to-left [29]. These models cannot look bidirectional at the same time. An NNLM that is bidirectionally trained can have a deeper sense of language context than unidirectional language models [29]. A recent model BERT (Bidirectional Transformers for Language Understanding) uses a bidirectional attention mechanism called Transformer to address this problem [29]. It uses a Masked LM that allows bidirectional training.

We fine-tune the pretrained BERT NNLM by adding a classification layer on top of the transformer layer for the downstream task of classification. The resultant classifier is used as the language-focused classification model in MAS. Figure 6 shows the BERT NNLM-based classification model used in MAS.

4 EMPIRICAL EVALUATION

The primary objective of experiments is to show the effectiveness of the MAS framework in instance selection for fine-tuning a transfer learning-based classification model. Given a budget (i.e., the number of instances to be labelled), we experimentally show that instances selected by MAS can better fine-tune a transfer learning-based classification model when compared with state-of-the-art active learning models.

4.1 The Transfer Learning-Based Classification Model for Downstream Tasks

In this research, we use BERT [29] as the NNLM transfer learning-based classification model. We individually fine-tune the pretrained BERT model using instances selected by each of the active
Active Learning for Effectively Fine-Tuning Transfer Learning to Downstream Task

Fig. 6. The LM-focused classifier based on the NNLM architecture with two transformer layers.

learning methods benchmarked. During fine-tuning, we add one additional output layer on top of the pretrained BERT model as shown in Figure 6. It has 24 hidden layers each with 1,024 hidden units, 16 self-attention heads, and 340M parameters. During unsupervised pretraining, BERT captures important general knowledge from the pretraining corpus (i.e., BooksCorpus with 800M words and Wikipedia with 2,500M words) and projects it to latent space by feature learning. During fine-tuning, BERT updates the initial latent space according to the dataset used in the downstream task. Therefore, a quality labelled dataset (i.e., representing diverse aspects) for fine-tuning the model effectively can improve the accuracy in a downstream task. In other words, the pretrained model can represent the general knowledge of a given text that helps in reducing model overfitting, and a fine-tuned model adds task-specific additional knowledge to the representation that improves model accuracy.

We evaluate the NNLM transfer learning-based classification models in two downstream tasks: (a) misogynistic tweet detection in two datasets, namely, Automatic Misogyny Identification (AMI) and QUT provided Misogyny Identification (QMI), and (b) hate tweet detection in the East Asia Hate (EAH) dataset.

4.2 Data Collection and Active Learning Settings

We use three datasets for comparing the effectiveness of MAS. Our assumed budget is 3,000 (i.e., our oracle can label a maximum of 3,000 tweets). In other words, our MAS model and the baseline active learning models are allowed to request for labelling a total of 3,000 tweets for a given dataset. The 3,000 labelled tweets selected by each model are used to fine-tune a transfer learning-based NNLM classification model. Classification performance achieved by each fine-tuning is evaluated using a held-out testing set and recorded to compare MAS and state-of-the-art active learning models.

4.2.1 AMI Dataset. The AMI English collection [33] contains a total of 4,000 and 1,000 labelled tweets in training and testing sets, respectively. The training set has 1,785 (45%) misogynous tweets and 2,215 (55%) non-misogynous tweets, while the testing set has 460 (46%) misogynous tweets and 540 (54%) non-misogynous tweets. Because the budget is 3,000, each of the state-of-the-art
models and the MAS model sample a set of 3,000 tweets from the training set for fine-tuning the transfer learning-based classification model.

4.2.2 EAH Dataset. The EAH dataset [77] contains East Asia hate tweets during COVID-19, collected between 1st January and 17th March 2020. EAH includes a total number of 20,000 tweets labelled as to whether a tweet is East Asian relevant and, if so, what the stance (Very Negative, Negative, Neutral, Positive, and Very Positive). A total of 3,898 instances are labelled positive (i.e., negative or negative stance towards East Asian people). To remove the skewness in the data class, we randomly selected a total of 3,898 instances labelled as Neutral, Positive, and Very Positive. This subset contains a total of 7,796 instances. For evaluating classifiers, we split this into 80%, 10%, and 10% for training, validation, and testing, respectively.

4.2.3 QMI Dataset. The QMI dataset was obtained using Twitter’s streaming API. A set of tweets that contain any of the three main misogynistic keywords (i.e., whore, slut, rape) was filtered. Tweets that contain a lot of non-English words were removed. A total of 5,000 tweets was obtained. Each tweet was then assessed with the following contextual information [11, 13] to label whether it contains targeted misogynistic abuse or not: (a) Is a specific person or group being targeted in this tweet? (b) Does this tweet contain a specific threat or wish for violence? (c) Does this tweet encourage or promote self-harm or suicide? (d) Is the tweet harassing a specific person, or inciting others to harass a specific person? (e) Does the tweet use misogynistic language in objectifying a person, making sexual advances, or sending sexually explicit material? (f) Is the tweet promoting hateful conduct by gender, sexual orientation, and so forth?

Out of 5,000 tweets, 1,800 (36%) tweets were found misogynistic and 3,200 (64%) were found non-misogynistic. The dataset is randomly divided into two subsets with 80% for training and 10% for testing. That is, the training and testing set contain 4,000 and 1,000 labelled tweets, respectively. As the labelling budget is 3,000, each of the state-of-the-art models and MAS samples/selects a set of 3,000 tweets from the training set for fine-tuning the transfer learning-based classification model.

4.3 Evaluation Measures

We used six standard classification evaluation measures [60]: Accuracy, Precision, Recall, F1 Score, Cohen Kappa Score (CKS), and Area Under the Curve (AUC). In addition, micro average (micro avg) and weighted average (weighted avg) were used to check that models are performing well in both minority and majority classes [71]. We also report True-Positive (TP), True-Negative (TN), False-Positive (FP), and False-Negative (FN) values.

4.4 Baseline Models

We use five state-of-the-art Active Learning models and a random selection model to compare MAS. Following is a brief description of the models.

—Learning Loss for Active Learning (LLAL) [83]: It is one of the latest methods that is task-agnostic, fast to train, and performs well on deep networks. It selects instances that a classification model is likely to produce a wrong prediction. A small parametric module is added to a neural network-based classifier to learn to predict the classifier’s losses on unlabelled inputs.

—MPDM [20]: It selects samples from instances in a batch such that the marginal probability distribution represented by the selected and labelled instances is closest to the distribution represented by the not-selected instances.
Table 2. Active Learning Performance of MAS on the AMI Dataset

|                  | MAS | HSAL | QBC | USM | MPDM | RSM | LLAL |
|------------------|-----|------|-----|-----|------|-----|------|
| True Positive    | 341 | 328  | 331 | 344 | 355  | 313 | 354  |
| True Negative    | 370 | 357  | 369 | 364 | 326  | 356 | 342  |
| False Positive   | 170 | 183  | 167 | 176 | 214  | 184 | 198  |
| False Negative   | 119 | 132  | 129 | 116 | 105  | 147 | 106  |
| Accuracy         | 0.711 | 0.685 | 0.704 | 0.708 | 0.681 | 0.669 | 0.696 |
| Precision        | 0.667 | 0.642 | 0.665 | 0.662 | 0.624 | 0.630 | 0.641 |
| Recall           | 0.741 | 0.713 | 0.720 | 0.748 | 0.772 | 0.680 | 0.770 |
| F1-Score         | 0.702 | 0.676 | 0.691 | 0.702 | 0.690 | 0.654 | 0.700 |
| CKS              | 0.423 | 0.371 | 0.408 | 0.418 | 0.369 | 0.338 | 0.397 |
| AUC              | 0.713 | 0.687 | 0.705 | 0.711 | 0.688 | 0.670 | 0.701 |

- HSAL [26]: It performs hierarchical clustering on the dataset in the batch mode and performs pruning of the cluster tree. This model has been commonly used to reduce sampling bias in active learning.
- Uncertainty Sampling Model (USM) [65]: It selects the instances for which the model is least certain how to label. It uses SVM [23] as the underlying model and selects the instances closest to the decision boundary. It is the most commonly used active learning framework [65].
- Query by Committee (QBC) [67]: A committee of models predicts the labels. Instances are selected according to the principle of maximal disagreement.
- Random Sampling Model (RSM): It randomly selects instances without replacement.

4.5 Results and Discussion

As the goal of this research is to sample/select instances that are relatively more effective than other instances in fine-tuning a transfer learning model for a downstream task. We fine-tune a classifier based on the language-based transfer learning model BERT [29] using each set of the labelled tweets selected by MAS and other state-of-the-art active learning models. For each method, we evaluate the classifier’s performance on the held-out test set.

4.5.1 Downstream Task 1: Misogynistic Detection. The evaluation results of all active learning methods on AMI and QMI datasets are given in Tables 2 and 4, respectively. Each column in the tables shows the classification performance when the transfer learning model BERT is fine-tuned with the set of labelled tweets selected by an active learning model.

The results on AMI dataset (Table 2) show that the NNLM classification model trained with the dataset obtained by MAS mostly outperforms models trained with datasets obtained by state-of-the-art active learning models. MAS improves accuracy, precision, recall, F1-score, CKS, and AUC of the classification model up to 6.28%, 5.96%, 8.95%, 7.37%, 25.27%, and 6.48%, respectively, when compared with the model trained with random sampling. These results ascertain that introducing a mixed aspect focused instance set can improve the fine-tuning performance of a transfer learning language model. It is to be noted that, using a simple ensemble of only two classifiers (one focusing linguistic aspects and the other focusing topical aspects), MAS achieves this current level of performance. Ensembling more classifiers focusing more aspects can provide better performance.

The AMI dataset was also used in Evalita 2018 classification competition [33]. Table 3 reports the results of methods participating in the competition. A comparison of Tables 2 and 3 shows fine-tuning BERT with only 3,000 training instances selected through MAS, QBC, or USM can achieve
Table 3. Results of Various Classification Models on the AMI Dataset in Evalita 2018 Competition [33]

| Rank | Team                | Accuracy |
|------|---------------------|----------|
| 1    | hateminers.c.run1   | 0.704    |
| 2    | hateminers.c.run3   | 0.681    |
| 3    | hateminers.c.run2   | 0.673    |
| 4    | resham.c.run3       | 0.651    |
| 5    | bakarov.c.run3      | 0.649    |
| ...  | ...                 | ...      |
| 18   | AMI-BASELINE        | 0.605    |
| ...  | ...                 | ...      |
| 22   | StopPropagHate.u.run3 | 0.591   |
| 23   | StopPropagHate.u.run2 | 0.59    |
| 24   | RCLN.c.run1         | 0.586    |
| 25   | SB.c.run3           | 0.584    |
| 26   | 14-exlab.c.run2     | 0.500    |

Table 4. Active Learning Performance of MAS on the QMI Dataset

|                | MAS | HSAL | QBC | USM | MPDP | RSM | LLAL |
|----------------|-----|------|-----|-----|------|-----|------|
| True Positive  | 301 | 292  | 297 | 286 | 280  | 278 | 291  |
| True Negative  | 552 | 548  | 539 | 559 | 569  | 555 | 557  |
| False Positive | 89  | 93   | 102 | 82  | 72   | 86  | 84   |
| False Negative | 61  | 70   | 65  | 76  | 82   | 84  | 71   |
| Accuracy       | 0.850| 0.837| 0.833| 0.842| 0.846| 0.831| 0.845|
| Precision      | 0.772| 0.758| 0.744| 0.777| 0.795| 0.764| 0.776|
| Recall         | 0.831| 0.807| 0.820| 0.790| 0.773| 0.768| 0.804|
| F1-Score       | 0.801| 0.782| 0.781| 0.784| 0.784| 0.766| 0.790|
| CKS            | 0.681| 0.653| 0.647| 0.660| 0.665| 0.633| 0.668|
| AUC            | 0.846| 0.831| 0.831| 0.831| 0.831| 0.817| 0.836|

better results than the top-performing team [33] that used all of the 4,000 training instances (their model was based on vector representation that concatenates sentence embedding, TF-IDF, and average word embeddings coupled with a Logistic Regression model). This comparison emphasises the importance of using MAS or some other active learning model for selecting instances to fine-tune transfer learning models.

To show that using MAS is generally useful, we used another dataset QMI. Results in Table 4 show that MAS outperforms state-of-the-art models. Results can further be improved by including other aspects besides linguistic and topical aspects in the MAS framework.

4.5.2 Ablation Study. The topic-focused classifier combines topical weight and polarity weight for estimating feature weight. Polarity weight is estimated based on the linear SVM, and topical weight is estimated based on topics discovered by LDA. For an ablation study of the topic-focused classifier, we compare it with a linear SVM using a small dataset. During the QMI data collection, an initial (or seed) set of 1,003 labeled tweets were collected manually. Information such as frequent words, patterns, and abuse targets was learned by analysing this dataset and was utilised during
Table 5. Performance Comparison of Linear SVM Classifier and Topic-Focused Classifier on a Subset of QMI Dataset

|                | Linear SVM |          |          |
|----------------|------------|----------|----------|
|                | Precision  | Recall   | F$_1$-score |
| Macro avg      | 0.69       | 0.67     | 0.67     |
| Weighted avg   | 0.70       | 0.71     | 0.70     |

|                | Topic-Focused Classifier |          |          |
|----------------|---------------------------|----------|----------|
|                | Precision | Recall | F$_1$-score |
| Macro avg      | 0.69 | 0.70 | 0.69 |
| Weighted avg   | 0.72 | 0.70 | 0.70 |

Table 6. Active Learning Performance on EAH Dataset

|                | MAS | LLAL | RSM |
|----------------|-----|------|-----|
| True Positive  | 343 | 331  | 325 |
| True Negative  | 313 | 319  | 317 |
| False Positive | 87  | 81   | 83  |
| False Negative | 37  | 49   | 55  |
| Accuracy       | 0.841 | 0.833 | 0.823 |
| Precision      | 0.798 | 0.803 | 0.797 |
| Recall         | 0.903 | 0.871 | 0.855 |
| F$_1$-Score     | 0.847 | 0.836 | 0.825 |
| CKS            | 0.683 | 0.667 | 0.647 |
| AUC            | 0.843 | 0.834 | 0.824 |

the subsequent QMI data collection. Table 5 shows the marginal performance improvement of the topic-focused classifier over the linear SVM when trained with this initial set of 1,003 QMI instances.

4.5.3 Downstream Task 2: Hate Abuse Detection. For further evaluation of the effectiveness of the proposed MAS model, we use the recently collected EAH dataset during COVID-19 [77]. Baseline models such as Hierarchical Sampling for Active Learning (HSAL), QBC, USM, and MPDP are very slow to train and select instances for labelling. Some of these models take days to train and select text data instances. Therefore, these models are difficult to deploy in practice. On the other hand, compared with these models, LLAL is fast, it is developed with special consideration of neural networks, and it performs well on deep networks [83]. Moreover, experimental results in Tables 2 and 4 show that LLAL performs better than other baseline models. RSM selects instances randomly, hence, it is very fast. Therefore, on the dataset EAH, we compare the MAS model against LLAL and RSM.

The experimental results in Table 6 show that the classifier model built with MAS data performs better than of LLAL and RSM in Accuracy, Recall, F$_1$-score, CKS, and AUC. LLAL has a marginally higher result in Precision. MAS has a higher Recall value that may provide higher false positives and lose some of the precision. In particular, the higher F$_1$-score of MAS indicates that MAS improves both precision and recall, not just precision, unlike LLAL. The authors in [77] used a total of 6,236 instances (i.e., 80% of 7,796 instances) to train several deep learning models to predict hate abuse. Empirical analysis shows that the best performance (F$_1$ of 0.83) was obtained with RoBERTa.
(BERT-like transfer learning model) and the F_1 score of 0.76 was obtained from the baseline LSTM. Our experimental results in Table 6 show that we achieve a better result with the BERT-based NNLM model by using 3,000 instances selected with MAS.

4.5.4 Discussion and Future Works. These experiments show that the MAS framework is useful for both the cases where the underlying data distribution is highly imbalanced and the underlying data distribution is balanced. The labelling process is time-consuming and it becomes challenging to collect a reasonably large quantity of labelled data, especially when the dataset is highly imbalanced. A systematic and autonomous method of sampling such as MAS is the way forward.

Specific to misogynistic or hate abuse detection tasks, MAS facilitates access to semantics and context information that is often not available (e.g., it is difficult to use dictionary-based semantics for the nature of noise in tweets (they do not follow a standard language sequence (format)) and difficult to know the context because of the small length of a tweet).

For the topic-focused classifier, in this research, we used LDA for learning topics from a tweet collection as LDA has a mathematical foundation in probability theory that makes it easy to explain our idea in the MAS framework. However, there are some other topic modelling techniques such as Non-negative Matrix Factorisation (NMF) that can be used. Topic modelling through those techniques might improve the effectiveness of MAS. Also, to keep MAS simple to explain, we only used topic-focused and language-focused aspects. Including other aspects such as semantic relations might improve MAS. The ensemble model in MAS checks if two classifiers agreed on labelling an instance to a specific class. However, training the ensemble model to learn to decide how much importance should be given to each classifier might also improve the framework. These are some improvements that we will consider in our future work.

5 CONCLUSION
This article presents a novel MAS framework to select the training instances that can most effectively fine-tune a transfer learning model to a downstream classification task, to maximise the utilisation of an instance-labelling budget (or minimise the labelling task of an oracle). The framework proposes to select instances that can cover multiple aspects in a dataset (e.g., topic, language) to complement and fine-tune the transfer learning model to the underlying data distribution of a target task. The MAS framework uses an ensemble of classification models, each focusing on a specific aspect in the dataset, which emphasises essential aspects of underlying data distribution. Moreover, a reduced number of instances need to be labelled because cluster-based down-sampling is applied to the three distinct groups of instances such as majority class, minority class, and uncertain.

A series of experiments were conducted to investigate the effectiveness of MAS on two classification tasks. In experiments, the set of labelled instances selected by MAS is used to fine-tune the classification model based on BERT (language model-based transfer learning). Similarly, the classification model is separately fine-tuned by each set of labelled instances selected by state-of-the-art active learning models. The experimental results show that MAS selects instances that can effectively fine-tune the transfer learning model to the downstream target task’s underlying data distribution. The classification task’s performance is improved when BERT is fine-tuned with MAS selected instances compared with those selected by state-of-the-art active learning models or randomly selected instances.

REFERENCES
[1] Resham Ahluwalia, Himani Soni, Edward Callow, Anderson Nascimento, and Martine De Cock. 2018. Detecting hate speech against women in english tweets. EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018), 194.
[2] Abdullah Semran Alharbi, Md Abul Bashar, and Yuefeng Li. 2018. Random-sets for dealing with uncertainties in relevance feature. In Australasian Joint Conference on Artificial Intelligence. Springer, 656–668.

[3] David Andrzejewski and David Buttsler. 2011. Latent topic feedback for information retrieval. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 600–608.

[4] Josh Attenberg and Foster Provost. 2010. Why label when you can search?: Alternatives to active learning for applying human resources to building classification models under extreme class imbalance. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 423–432.

[5] Josh Attenberg and Foster Provost. 2011. Inactive learning?: Difficulties employing active learning in practice. ACM SIGKDD Explorations Newsletter 12, 2 (2011), 36–41.

[6] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 759–760.

[7] Md Abul Bashar and Yuefeng Li. 2017. Random set to interpret topic models in terms of ontology concepts. In Australasian Joint Conference on Artificial Intelligence. Springer, 237–249.

[8] Md Abul Bashar and Yuefeng Li. 2018. Interpretation of text patterns. Data Mining and Knowledge Discovery 32, 4 (2018), 849–884.

[9] Md Abul Bashar, Yuefeng Li, and Yang Gao. 2016. A framework for automatic personalised ontology learning. In IEEE/WIC/ACM International Conference on Web Intelligence (WI’16). IEEE, 105–112.

[10] Md Abul Bashar, Yuefeng Li, Yan Shen, Yang Gao, and Wei Huang. 2017. Conceptual annotation of text patterns. Computational Intelligence 33, 4 (2017), 948–979.

[11] Md Abul Bashar, Richi Nayak, and Nicolas Suzor. 2020. Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set. Knowledge and Information Systems (2020), 1–26.

[12] Md Abul Bashar, Richi Nayak, Nicolas Suzor, and Bridget Weir. 2018. Misogynistic tweet detection: Modelling CNN with small datasets. In Proceedings of the 16th Australasian Data Mining Conference.

[13] Md Abul Bashar, Richi Nayak, Nicolas Suzor, and Bridget Weir. 2018. Misogynistic tweet detection: Modelling CNN with small datasets. In Australasian Conference on Data Mining, Springer, 3–16.

[14] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. The Journal of Machine Learning Research 3 (2003), 993–1022.

[15] Guillaume Bouchard and Bill Triggs. 2004. The tradeoff between generative and discriminative classifiers. In Proceedings of the 16th IASC International Symposium on Computational Statistics (COMPSTAT’04), 721–728.

[16] Mausam C. Lin. 2018. Active learning with unbalanced classes and example-generated queries. In Proceedings of the AAAI Conference on Human Computation.

[17] Deng Cai and Xiaofei He. 2012. Manifold adaptive experimental design for text categorization. IEEE Transactions on Knowledge and Data Engineering 24, 4 (2012), 707–719.

[18] Wenbin Cai, Muhan Zhang, and Ya Zhang. 2015. Active learning for ranking with sample density. Information Retrieval Journal 18, 2 (2015), 123–144.

[19] Chris K. Carter and Robert Kohn. 1994. On Gibbs sampling for state space models. Biometrika 81, 3 (1994), 541–553.

[20] Rita Chattopadhyay, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. 2013. Batch mode active sampling based on marginal probability distribution matching. ACM Transactions on Knowledge Discovery from Data (TKDD) 7, 3 (2013), 13.

[21] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005.

[22] Chaitanya Chemudugunta, America Holloway, Padhraic Smyth, and Mark Steyvers. 2008. Modeling documents by combining semantic concepts with unsupervised statistical learning. The Semantic Web-ISWC 2008 (2008), 229–244.

[23] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. Machine Learning 20, 3 (1995), 273–297.

[24] Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In JSMB, Vol. 1999, 77–86.

[25] Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In Advances in Neural Information Processing Systems. 3079–3087.

[26] Sanjoy Dasgupta and Daniel Hsu. 2008. Hierarchical sampling for active learning. In Proceedings of the 25th International Conference on Machine Learning. ACM, 208–215.

[27] Thomas Davidson, Dana Warmley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv:1703.04009.

[28] William H. E. Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. Journal of Classification 1, 1 (1984), 7–24.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[30] Allen Downey. 2012. Think Bayes: Bayesian Statistics Made Simple. Green Tea Press.

[31] Henry R. Ehrenberg, Jaeho Shin, Alexander J. Ratner, Jason A. Fries, and Christopher Ré. 2016. Data programming with DDLite: Putting humans in a different part of the loop. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics. ACM, 13.

[32] Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. 2007. Learning on the border: Active learning in imbalanced data classification. In Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management. ACM, 127–136.

[33] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 task on automatic misogyny identification (AMI). Proceedings of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA’18).

[34] Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In Advances in Neural Information Processing Systems. 1019–1027.

[35] Yang Gao, Yuefeng Li, Raymond Y. K. Lau, Yue Xu, and Md Abul Bashar. 2017. Finding semantically valid and relevant topics by association-based topic selection model. ACM Transactions on Intelligent Systems and Technology (TIST) 9, 1 (2017), 1–22.

[36] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering 10, 4 (2015), 215–230.

[37] John Goutsias, Ronald P. S. Mahler, and Hung T. Nguyen. 2012. Random Sets: Theory and Applications. Vol. 97. Springer Science & Business Media, Berlin, Germany.

[38] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. arXiv preprint arXiv:1802.06210.

[39] Marti A. Hearst, Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. IEEE Intelligent Systems and Their Applications 13, 4 (1998), 18–28.

[40] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation 9, 8 (1997), 1735–1780.

[41] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1. 328–339.

[42] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. 2010. Active learning by querying informative and representative examples. In Advances in Neural Information Processing Systems. 892–900.

[43] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised graph-based topic labelling using DBpedia. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining. ACM, 465–474.

[44] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. arXiv preprint arXiv:1605.08151.

[45] Rudolf Kruse, Erhard Schwecke, and Jochen Heinsohn. 2012. Uncertainty and Vagueness in Knowledge Based Systems: Numerical Methods. Springer Science & Business Media, Berlin, Germany.

[46] David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In SIGIR’94. Springer, 3–12.

[47] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. arXiv preprint arXiv:1611.01368.

[48] Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In Proceedings of the 13th International Workshop on Semantic Evaluation. 87–91.

[49] Wenhe Liu, Xiaojun Chang, Yan Yan, Yi Yang, and Alexander G. Hauptmann. 2018. Few-shot text and image classification via analogical transfer learning. ACM Transactions on Intelligent Systems and Technology (TIST) 9, 6 (2018), 1–20.

[50] Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. arXiv preprint arXiv:1803.02893.

[51] Gábor Melis, Chris Dyer, and Phil Blunsom. 2017. On the state of the art of evaluation in neural language models. arXiv preprint arXiv:1707.05589.

[52] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing LSTM language models. arXiv preprint arXiv:1708.02182.

[53] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Proceedings of the 11th Annual Conference of the International Speech Communication Association.
[55] Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *IEEE Spoken Language Technology Workshop (SLT’12)*. IEEE, 234–239.

[56] Ilya Molchanov. 2006. *Theory of Random Sets*. Springer.

[57] Ilya S. Molchanov et al. 2005. *Theory of Random Sets*. Vol. 19. Springer.

[58] Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

[59] David Martin Powers. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology* 2, 1 (2011), 37–63.

[60] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

[61] Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. *Improving Language Understanding with Unsupervised Learning*. Technical Report. OpenAI.

[62] G. B. Ran, A. Navot, and N. Tishby. 2004. Kernel query by committee (KQBC). *Bachrach* (2004).

[63] Oscar Reyes and Sebastián Ventura. 2018. Evolutionary strategy to perform batch-mode active learning on multi-label data. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9, 4 (2018), 1–26.

[64] H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*. ACM, 287–294.

[65] Manali Sharma and Mustafa Bilgic. 2017. Evidence-based uncertainty sampling for active learning. *Data Mining and Knowledge Discovery* 31, 1 (2017), 164–202.

[66] Lixin Shi, Yuhang Zhao, and Jie Tang. 2012. Batch mode active learning for networked data. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 2 (2012), 1–25.

[67] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[68] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*. 843–852.

[69] Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research* 2 (Nov. 2001), 45–66.
[83] Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 93–102.

[84] Xinge You, Ruxin Wang, and Dacheng Tao. 2014. Diverse expected gradient active learning for relative attributes. IEEE Transactions on Image Processing 23, 7 (2014), 3203–3217.

[85] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero. 2010. Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. Computer Speech & Language 24, 3 (2010), 433–444.

[86] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329.

[87] Cha Zhang and Tsuhan Chen. 2002. An active learning framework for content-based information retrieval. IEEE Transactions on Multimedia 4, 2 (2002), 260–268.

[88] Kelly W. Zhang and Samuel R. Bowman. 2018. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. arXiv preprint arXiv:1809.10040.

[89] Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? The challenging case of long tail on Twitter. Semantic Web 10, 5 (2019), 925–945.

[90] Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K. Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 1137–1144.

Received July 2020; revised November 2020; accepted December 2020