Received: 1 February 2023 | Revised: 27 March 2023 | Accepted: 19 April 2023

DOI: 10.1111/exsy.13329

ORIGINAL ARTICLE

Expert Systems WILEY

Enhancing Arabic-text feature extraction utilizing label-semantic augmentation in few/zero-shot learning

Seham Basabain[1,2] | Erik Cambria[3] | Khalid Alomar[1] | Amir Hussain[2]

[1]Faculty of Computing and Information Technology, King AbdulAziz University, Jeddah, Saudi Arabia

[2]School of Computing, Edinburgh Napier University, Edinburgh, UK

[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

Correspondence
Seham Basabain, Faculty of Computing and Information Technology, King AbdulAziz University, Jeddah, Saudi Arabia.
Email: ssbasabain@kau.edu.sa; seham.basabain@napier.ac.uk

Funding information
U.K. Engineering, and Physical Sciences Research Council (EPSRC), Grant/Award Numbers: EP/T024917/1, EP/T021063/1, EP/M026981/1

Abstract

A growing amount of research use pre-trained language models to address few/zero-shot text classification problems. Most of these studies neglect the semantic information hidden implicitly beneath the natural language names of class labels and develop a meta learner from the input texts solely. In this work, we demonstrate how label information can be utilized to extract enhanced feature representation of the input text from a Transformer-based pre-trained language model such as AraBERT. In addition, how this approach can improve performance when the data resources are scarce like in the Arabic language and the input text is short with little semantic information as is the case using tweets. The work also applies zero-shot text classification to predict new classes with no training examples across different domains including sarcasm detection and sentiment analysis using the information in the last layer of a trained classifier in a transfer learning setting. Experiments show that our approach has a better performance for the few-shot sentiment classification compared to baseline models and models trained without augmenting label information. Moreover, the zero-shot implementation achieved an accuracy up to 0.874 in Arabic sarcasm detection from a model trained on a sentiment analysis task.

KEYWORDS
Arabic text classification, contextual embeddings, feature extraction, few/zero-shot learning, label semantics

1 | INTRODUCTION

Text classification problems have different approaches, some of these approaches utilize architectures with on-top linear layers trained to hold information to output class distribution for a given classification task. These approaches achieved state-of-the-art results and proved their effectiveness, but still suffer from some limitations in which, the number of classes must be predefined prior to training. Moreover, when training a model, the entire model is fine-tuned using the whole training samples of all predefined classes (Halder et al., 2020). This approach can provide superior performance when the classification task has classes with a reasonable amount of training examples.

In real-world scenarios, this might not be applicable for all text classification problems, where some scenarios are lacking annotated corpora or changing data over rapid times. These challenges motivate researchers to study methods in which text classifiers can be trained for new unseen classes with few training samples available (Halder et al., 2020).

Human learning systems have the advantage of learning new concepts from few support samples, this initiates a gap in computer learning systems when trained using scarce training data and motivates researchers to pay more attention to few-shot learning studies through the past

decade (Luo et al., 2021). Machine learning has succeeded in applications with a huge amount of data but is inhibited when the data set is very small. Few-shot learning can solve this problem in machine learning by applying meta learning techniques.

Meta learning is a type of machine learning whereby learning algorithms are automated and applied to metadata. It helps in observing how different artificial intelligence proceeds to perform some tasks based on 'learning to learn' (Vanschoren, 2018). Additionally, meta learning provides a paradigm that helps machine learning gain experiences through many episodes of lessons toward model improvement.

On top of that, meta learning is being used in different fields, such as transfer learning and hyperparameter optimization. Which follows a standard approach uses the experiences from an existing model for classifying task *X* to help improve learning and model weights initialization for the new classification task *Y*. Second, it is also being used in the adaptation and generation of the domain.

Domain adaptation is when the transfer learning variant tries to ease the domain shift by adopting a source-trained model (Wang et al., 2021). Lastly, it is being used in automated machine learning, that helps automate parts in machine learning. However, this approach disregards two major information since the source task differs from the target task. In this case, the final output linear layer acting as a model decoder is dismissed with all its information and a new decoder must be trained from scratch. Another major weakness of this approach is the dismissal of the semantic information beneath class labels (Halder et al., 2020). These two sources of information can be useful and sufficient to train new classifiers with few or even no training data as in few-shot or zero-shot learning scenarios.

Appending class names to an input sentence will help humans to interpret the semantics of a sentence for a given class. This also would be effective when applying this method to any text classification task (Halder et al., 2020), where the goal is to find a mapping function *f* as:

$$f : text \rightarrow \{0,1,2\}^{M} \text{ that is, } f(t) = P(t \mid y_s) \forall s \in \{1 \ldots M\} \quad (1)$$

where *t* refers to the input text mapped to an *M*-dimensional vector, each label (y_s) whether it is available or not is denoted by probability *p* and identified by a dimension *s*.

Given an example of the above formulation, we can factorize a text classification problem as a learning function from the input text. We can leverage the semantic information beneath these text inputs by adding class names to the function and checking the extent of matching between the two inputs (Halder et al., 2020). In other words, for a sentiment classification problem trained to predict the polarity of a piece of text, we can create a tuple that consists of both the sentence with its label definition as:

<"مشاعر إيجابية","أنا أحب أطفالي">

Where the output being either positive, negative or neutral based on the augmented information of the label semantics. In this case, the sentence 'أنا أحب أطفالي' which means 'I love my kids', explicitly shows a positive polarity. Appending label semantic information 'مشاعر إيجابية' which is 'positive sentiment' can interpret semantics beneath the text. This would be more effective when a text input implicitly carries semantics of sentiment classes such as 'وائل رجال يوزن بالذهب' meaning 'Wael is a man whom worth his weight in gold', this statement contains a metaphor with implicit positive semantics to imply that this man is valuable. Adding label information to this input can help a classifier to predict the class of a training example leveraging the semantic information contained in a trained linear layer.

To this end, this research aims to contribute to the field of Arabic text classification the following:

1. To analyse whether adding label semantic information in a few/zero-shot setting will strengthen the capability of a Transformer-based model such as AraBERT to extract leveraged representations of short Arabic text.
2. To build a prototypical network meta learner and augment this meta learner with label semantics.
3. To build a generalized meta learner framework able to classify short Arabic text in cross-domain tasks by generating a general decoder for different classification tasks.
4. To use this generalized framework for text classification of short Arabic text in a zero-shot learning setting.

The study then continues to provide a brief representation of related work in Section 2. Then Section 3 illustrates details of the materials and methodology. Next, the experimental setup is illustrated in Section 4. Then, Section 5 discusses experimental results to evaluate the proposed approach on the selected data sets. Finally, the paper's conclusion is presented in Section 6.

2 | RELATED WORK

This section surveys literature to review what few-shot learning is, what techniques are available to implement this approach and related work implementing text classification tasks in few/zero-shot learning.

2.1 | Few-shot learning

Few-shot learning is the process of classification or regression based on a very small number of samples. It somehow mimics how human learning systems can distinguish any new objects only based on very few samples learned before. When applying deep learning models for standard supervised learning, it is hard to train a model using only a few or no samples of training data. However, the goal of few-shot learning is not to train a deep learning model to recognize samples in a training set to generalize on the test set. Instead, it trains the model to make it able to 'learn to learn' how to distinguish different inputs based on similarities and differences between them rather than 'learn to generalize' such as in the standard supervised learning (Wang, 2022).

Training data in a few-shot setting unlike the standard supervised learning does not have to include samples for all classes of prediction. For example, a model can be trained on different image samples of classes Tiger, Rabbit and Monkey, then new images of cats that the model has not seen before are fed to the model. In this case, the model has not trained on cats during training, thus it is not able to generalize that these images belong to the Cat class. However, since a few-shot setting is to train the model on the concept of making it able to learn differences and similarities, it can tell that these images of cats are similar and belong to the same object.

The model compares the samples in a query set with each sample in different classes in the support set. *Support set* is a meta learning terminology, it is a small set of limited classes holding few samples. It is different from regular training sets used to train deep neural networks, in which, these small support sets provide additional information to the model during test time even though these classes provided in the support set were not available in the big training set during model training. *Query set* is a set of samples belonging to classes not seen before during the training process; these unknown classes are not among the classes in the training set. In this case, the model does not know which class this sample belongs to, thus, the support set during training provides more information about the specifics of samples in the query set. Samples in the query set can be compared to those in the support set to recognize similarities and differences.

Terminologies in few-shot learning include the *k-way n-shot* support set. This means that during training the model is provided with a support set that consists of *k* classes, each class has *n* samples. For example, a 3-way 2-shot support set would include two samples in each of the three classes, Table 1 gives a possible example.

The number of ways *k* in the support set can affect the prediction accuracy of a model, when *k* is increased the accuracy of a model prediction decreases, unlike the number of shots per class when this number is increased, the prediction accuracy improves (Wang, 2022). This is due to the ability of a model to learn a similarity function between a query sample with every sample in the support set based on multiple examples in a few sets of classes rather than multiple unrelated classes.

2.1.1 | Few-shot learning methods

To tackle the problem of data scarcity, few-shot learning can make use of a small set of data with supervised information providing some prior knowledge to predict a target input (Wang et al., 2021). Few-shot learning methods depending on how prior knowledge is utilized can be classified into three groups: *information* that uses the previous comprehension to expand on the superintendent experience. Secondly, the *model* that uses previous knowledge to decrease the dimension of the room in the hypothesis, used when the complexity of the hypothesis is being constrained, which leads to a minimal hypothesis. Lastly, the *technique* uses previous knowledge to change the search for a hypothesis, which is the algorithm used in few-shot learning to find sign theta in an equation (Choi et al., 2018). Where sign theta is a hypothesis variable that is crucial in zero-shot learning. The following sections illustrate some of these methods.

Transfer learning approach

Text classification scenarios in the real world usually encounter data deficiency problems (Geng et al., 2019). To overcome this issue, some methods allow training classifiers with only little training examples. Transfer learning is one of these methods, in which it is a process whereby knowledge is transferred from the original domain with much instruction and data used for training is in abundance to a target domain where

TABLE 1 Example of 3-way 2-shot support set in few-shot learning.

Positive	Negative	Neutral
I love eating Pizza	The movie is awful	His name is Mohammed
The weather is amazing	I hate physics	Today is March 26th

2-shot

3-way

instruction data is little. Domain adaptation is an example of transfer learning. In domain adaptation, the task of the sources is the same, while the domain targets are entirely different. In this case, the whole trained model for the source task can be disregarded and trained from scratch for the target task. However, if the two tasks are somehow similar, then the fine-tuned encoder of the source model can be transferred to the target task (Halder et al., 2020).

An example of this is when the source task contains customers' reviews about a movie and the target domain contains the purchases which customers made to acquire a movie. The few-shot learning typically uses the transfer learning method whereby the previous knowledge acquired is conveyed from the source to the task of the low-shot.

Metric learning approach

This approach is basically to train a model how to 'learn to compare' through learning a distance metric. There are some standard methods in metric learning approaches such as: induction networks, prototypical networks, relation networks and matching networks.

The **Induction networks** propose the introduction of an induction module capable of inducing the prototype using dynamic routing (Geng et al., 2019). The network aims to improve the prototype reduction using routing that is dynamic and algorithms calculated from the capsule network to take the place of the operational mean. The representation on class levels is derived dynamically using LSTMs as the encoder in a manner that is not parametric (Tang et al., 2020).

The **Prototypical networks** have been introduced by (Snell et al., 2017), these networks rely on partitioning a number of N data points into K prototypes or clusters, in which each cluster holds points with the nearest mean. This method is easy to apply and commonly used in research due to its simplistic and inductive bias which is useful in the regime of limited data to achieve exceptional results (Choudhury, 2021).

Figure 1 shows that the network consists of 3-way 5-shot, for each prototype a centroid (c_1, c_2 and c_3) which is an average value of each data point vector initiated through neural network encoding. Once all support set samples are mathematically represented and vectors are averaged to form a class prototype, a new input X from the query set is embedded through the embedding space, to decide which prototype this new input vector belongs to, a Euclidean distance is calculated between this point and all three prototypes c_1, c_2 and c_3 as denoted by the dotted lines. The higher probability value $-d(X, c_k)$ means a closer distance between the two, which means X belongs to the c_2 prototype.

The **Matching network** was the first to be developed in solving few-shot learning problems. This network is complex and it compares items in the query set to those in the support set, relying on embedding query sets into a sequence of support sets via BiLSTM. As shown in Figure 2, a large-base data set is used to solve the task of few-shot learning. It computes input embeddings with no prior knowledge of the classes. The process takes place through the comparison of different class instances for all the classes. Due to the difference in the classes for every episode, the network extracts input features relevant to the discrimination between the classes. The algorithm studies the features specific to each class in the case of standard classifications. Improvements have been made to the original algorithm like the dependence of one input embedding on the embedding of other inputs (Vinyals et al., 2016).

Both the matching and the prototypical networks were established for the mitigation of the shortcomings brought by the few-shot learning technique. The difference between the two is that the matching networks produce a classifier to the weighted nearest neighbour while the prototypical networks produce a classifier that is linear using the Euclidean distance.

Another method of the metric learning approach is **Relation networks**, in this method, the relation module is exploited to model the relationship between vectors of the query set and class-wise vectors of the support set (Sung et al., 2018). These two feature maps extracted from each input capturing specific information are then concatenated and a relation score between 0 and 1 is obtained through weight optimization g_ϕ. This relation score represents the similarity between the features of the support set and the features of the query set to classify which class the query input belongs to. Figure 3 illustrates the architecture of this network.

Augmentation approach

Methods used in few-shot learning use prior knowledge so that the supervised information can make sense. When the sample set is augmented, the data is enough to get a hypothesis that is reliable and true. Samples transformation entailing the training data entails the augmented strategy

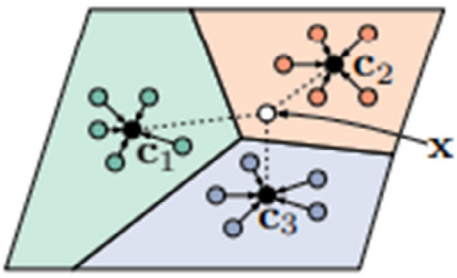

FIGURE 1 Prototypical networks in few-shot learning.

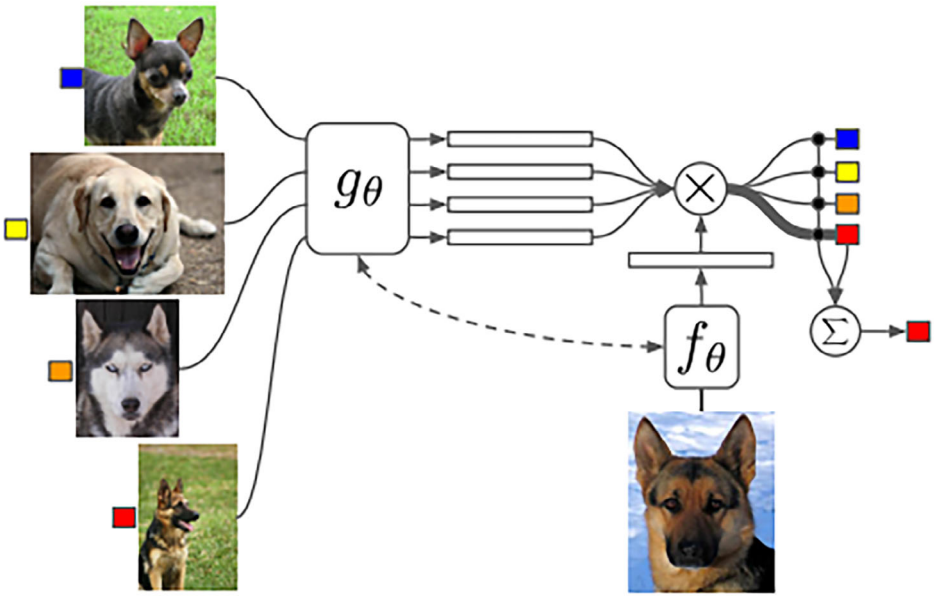

FIGURE 2 Matching network in few-shot learning.

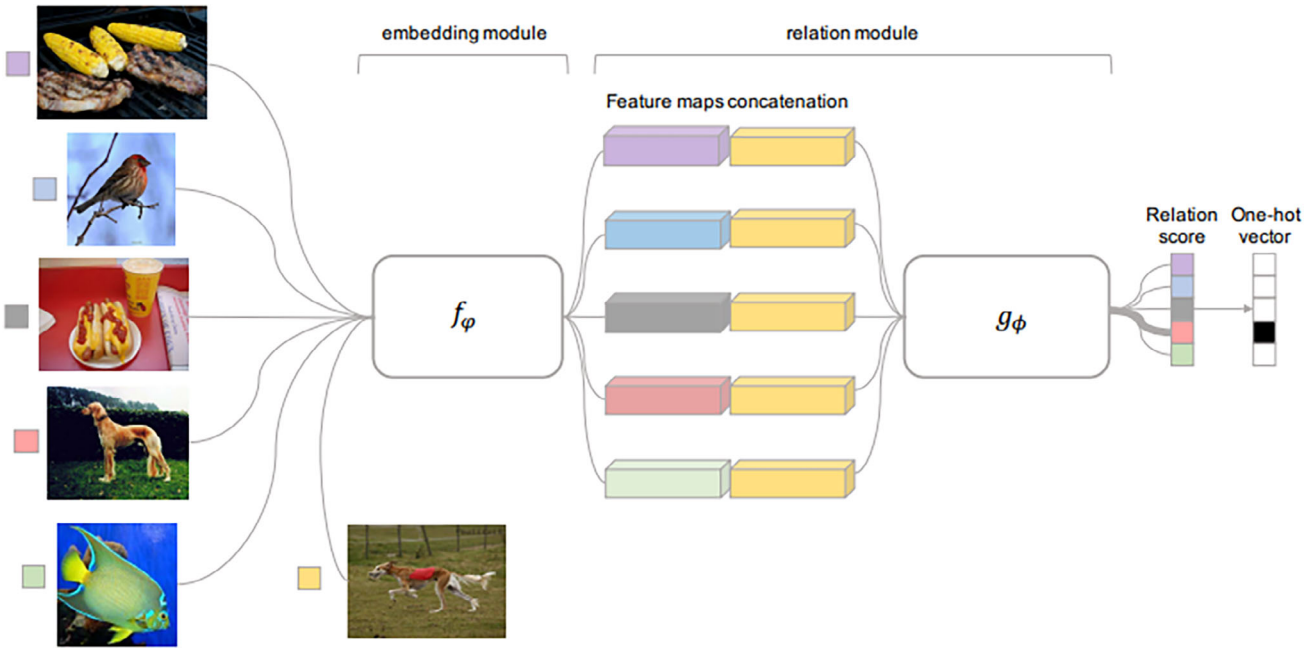

FIGURE 3 Relation network in few-shot learning.

that transforms the samples into specific variations. The procedure of transformation includes experience as previous knowledge, which helps get some other information (Sung et al., 2018).

The augmentation approach trains a model how to be able to 'learn to augment', utilizing a generator which can learn from a support set to generate new data, this can be done using a source domain data and a discriminator generating other similar data within class data samples (Antoniou et al., 2017).

2.2 | Few-shot and zero-shot text classification

Surveying literature, few-shot learning approaches have been utilized in various research works to solve different text classification tasks. The study of Hofer et al. (2018), aimed to improve the execution of the named entity recognition (NER) task of i2b2 2009 while using just

10 annotated discharge summaries which were chosen at random. The authors utilized six medical NER data sets, distributed as follows: one which determines the target task and supervised training and testing, two for supervised pre-training of weights, and three for unsupervised training of custom word embeddings. As well as they utilized a nonmedical data set for supervised pre-training of weights. Each of the data sets used in a supervised fashion offered several target NER categories to be used as labels. However, the study assessed the layer-wise initialization with pre-trained weights, hyperparameter tuning, combining pre-training data, custom word embeddings, and optimizing out-of-vocabulary (OOV) words. The major findings from the experimental results revealed that the F1 score of 69.3% achievable by state-of-the-art models can be improved to 78.87%.

On the other hand, the study of Chen et al. (2019), discussed the ability to notably decrease human annotation effort to achieve logical performance through neural natural language generation models and discussed the supposed processes that provide a top performance of generative pre-training, to develop text from a structured data. As per these issues, the authors proposed a model which is based on content selection from input data and language modelling to form coherent sentences, in which it is obtained from previous knowledge. The study conducted 200 training examples, from different domains. However, the major results showed that this proposed model attained highly rational performances and surpassed the most vigorous baseline by an average of over 8.0 BLEU points improvement.

Moreover, the study of Chada and Natarajan (2021), discussed the significant task of learning from only a few examples and its relevance to a real-world setting. The study argued the question-answering, as the existing state-of-the-art pre-trained models mostly require fine-tuning on huge numbers of examples to acquire satisfying outputs, and how their performance decreases notably in a few-shot setting to less than 100 examples. The authors tried in their research to tackle this issue by introducing a fine-tuning framework that influences the pre-trained text-to-text models and stands straight with its pre-training framework. The authors create an input as a series of the question, a mask token representing the answer span and a context. However, from major findings of the experimental results, it was revealed that this formulation drives important gains on multiple question-answering benchmarks. These gains increase as larger models are utilized and translate well to a multilingual setting.

Another study by Gu et al. (2021), argued prompts for pre-trained language models and their notable execution by covering the gap amidst pre-training tasks and other different tasks. The study stated that within these methods, prompt tuning, which freezes pre-trained language models and only tunes soft prompts, offers a functional resolution for adapting large-scale pre-trained language models to downstream tasks. The authors added that prompt tuning is not yet discovered precisely. The study aimed to identify the utilization of pre-trained language models for few-shot learning dynamically through prompt tuning. For this purpose, the authors executed pilot experiments to interpret the performance of prompt tuning on pre-trained language models. However, the current research found that prompt tuning surpasses other prompt tuning baselines, performing comparable to or even better than full-model tuning.

The study of Wu et al. (2021), argued that GPT-3 had an effective zero-shot and few-shot learning on several natural language processing (NLP) tasks by scaling up the model size, data set size and computation size. Yet, training such a model needs a big quantity of computational resources which is a challenging issue for scholars. For this purpose, the study introduced a procedure that combines large-scale distributed training performance into a model architecture design. This method is called Yuan 1.0 which is a large-scale pre-trained language model in zero-shot and few-shot learning. Major findings showed that the proposed solution offers a profound capacity for NLP tasks and has articles that are hard to be differentiated from the articles written by humans.

Moreover, the study of Xu et al. (2022), aimed to propose the CP-Tuning, which represents the first end-to-end Contrastive Prompt Tuning framework for fine-tuning pre-trained language models without any manual engineering of task-specific prompts and verbalizers. This framework is consolidated with the task-invariant persistent prompt encoding technique and has totally trained prompt parameters. The study also introduced the pair-wise cost-sensitive contrastive learning process to enhance the model and make it able to fulfil verbalizer-free class mapping and enhance the task-invariance of prompts. It learns to differentiate various classes and makes the decision boundary easier by allocating various costs to simple and tough cases. However, major empirical results showed that among a collection of language understanding tasks that had been utilized in information retrieval systems and various pre-trained language models, the proposed CP-Tuning surpasses the state-of-the-art methods.

In addition, the study of Xia et al. (2022), aimed to prove that ELECTRA surpasses other masked language models in a variety of different tasks. However, the authors argued that ELECTRA is a profound effective substitutional in the full-shot setting, which executes few-shot learning by creating downstream tasks as text infilling. Major experimental results showed that ELECTRA learns distributions that stand superior with downstream tasks.

Another study by Xu et al. (2021), aimed to propose a new knowledge-enhanced pre-trained model named KEBERT, which realizes the unstructured knowledge from enormous text corpora and the structured knowledge from knowledge graphs. The study also aimed to introduce a few-shot learning algorithm named Fuzzy-PET to enhance the generalization capabilities of pre-trained language models for few-shot learning. Furthermore, the scholars tried to reach the top effectiveness in limited and unlimited tracks of FewCLUE with their proposed solution, that is going to be offered to the public. However, the introduced resolution FewCLUE is a novel benchmark of few-shot learning for Chinese language understanding evaluation. It has nine challenging tasks involving an inclusive language understanding subjects with text classification, language inference, idiom comprehension and co-reference resolution. Tasks are distributed as every task includes 5 independent labelled subsets, each

subset with 16 examples in a class for training and the same amount of data for validation, and extra unlabelled examples. Models had been assessed in accordance with the averaged test performance trained over five subsets per task. The major experimental findings revealed that the generated models achieve the best performance in both limited and unlimited tracks of FewCLUE.

For zero-shot text classification, the study of Ma et al. (2016), discussed the challenges faced by the fine-grained named entity typing or what is known as (FNET). As per the authors, the challenges are two main issues: the growing type set and label noises, as the type of hierarchy of organizations is usually built from knowledge bases for example DBpedia, this knowledge base is orderly updated with novel types particularly the fine-grained ones, and organizations, as it is normal to suppose that the type of hierarchy is growing instead of being stable over time. Yet, the authors stated that the actual fine-grained named entity typing systems are blocked from treating with a growing type set for that information learned from the training set cannot be transferred to unseen types. The second issue with fine-grained named entity typing is the weakly supervised tagging procedure, that is utilized for unprompted produced labelled data and automatically inserts label noises. However, the authors proposed a novel technique which combines prototypical and hierarchical information to learn pre-trained label embeddings. They also conform to a zero-shot framework which can foresee both seen and prior unseen entity types. The study then implemented assessment on three benchmark data sets with two settings: the first one is few-shot recognition in which the whole types are wrapped by the training set and the second is zero-shot recognition where fine-grained types are supposedly missing from the training set. Moreover, major findings of this research article showed that previous knowledge encoded utilized by the current proposed label embedding technique can considerably raise the effectiveness of classification in both scenarios.

On the other hand, the study of Roy et al. (2022) discussed the zero-shot learning major issue of training and testing on a totally disjoint set of classes, in which they explained that the problem exists in its capacity to transmit knowledge from train classes to test classes. They added that classical semantic embeddings consisting of human defined attributes (HA) or distributed word embeddings (DWE) are utilized in order to simplify this transfer by enhancing the linkage between visual and semantic embeddings. This research article tried to benefit from the evident linkages between nodes defined in ConceptNet (Speer et al., 2017), a common-sense knowledge graph, to produce common sense embeddings of the class labels by utilizing a graph convolution network-based autoencoder. Their major findings showed that the experiments on three standard benchmark data sets exceeded the strong baselines when they integrated their common-sense embeddings with existing semantic embeddings such as HA and DWE.

Few-shot learning studies on the Arabic language are still in their infancy, this is due to the scarcity of Arabic linguistic resources, corpora and lexicons (Alwaneen et al., 2022). The study of Hardalov et al. (2022), had discussed the purpose of using the stance detection, where the authors stated that this technique is used to obtain the context (viewpoint) identified in a text across a target. Such viewpoints are predominantly identified in several distinct languages based on the user and on the platform. The research added that several studies on stance detection were restricted to a single language and on a small number of targets, and with limited utilization of cross-lingual stance detection. Furthermore, the issue is that non-English sources of labelled data are scarce, in which extra complexities are being faced. However, the authors mentioned that huge multilingual language models have strongly enhanced the effectiveness of a big number of non-English tasks, particularly those with a limited quantity of examples like the Arabic language. Consequently, the study focuses on the significance of pre-training of models and how it can learn from a few examples. To prove this, the authors tested 15 various data sets in 12 languages from six language groups, as well as tested 6 low-resource assessment settings. Major findings from this article showed that there was a notable enhancement of greater than 6% F1 absolute in few-shot learning settings in comparison with several profound baselines.

Moreover, the study of Khalifa et al. (2021), had argued the issue of the high cost associated with achieving labelled data, particularly in the case of numerous different languages and accents. The authors stated that fine-tuning of the pre-trained language models for downstream tasks needs an adequate quantity of data being annotated. For this purpose, the authors suggest a model that has a self-train pre-trained language with zero-shot and few-shot scenarios to enhance the effectiveness on data-rareness diversities based on resources from data-rich ones. The study explained the usage of its method in the context of Arabic sequence labelling through language model fine-tuned on modern standard Arabic (MSA) just to foresee named entities (NE) and part-of-speech (POS) tags on several dialectal Arabic (DA) assortments. However, the major findings of this research study showed that self-training is a vigorous method, in which the zero-shot MSA-to-DA transfer is enhanced by ∼10% F1 NER and 2% accuracy POS tagging.

2.3 | Text classification using label semantic information

Few-shot learning has multiple methodologies, meta learning is considered as one of the dominant approaches that relies on a function that maps a few support samples to a classifier using a meta training data set (Finn et al., 2017). Few-shot learning has the capability of learning new concepts from a few training samples, meta learning can be utilized then to extract class-relevant features from an input text in the support set to be most compatible with the query features.

Meta learners and pre-trained language models proved an outstanding performance in many existing systems with a limited number of training data (Luo et al., 2021). However, although semantics can play a major role in providing more information and reducing the ambiguity of the

input text with different intent classes, these existing systems do not utilize semantics of class labels, instead, they build these meta learners using the information of the input text solely (Luo et al., 2021).

Moreover, despite that meta learning frameworks and pre-trained language models performed very well in few-shot problems, they were mostly used in computer vision (Luo et al., 2021). Few-shot learning in the field of NLP and text classification has recently been introduced and proved an outstanding performance when little data is transferred to downstream tasks.

In the field of text classification, few-shot learning has been introduced in some recent works through past years. ROBUSTTC-FSL applies an approach of adaptively selecting a metric learning for an optimal distance metric of multiple tasks (Yu et al., 2018). Another work by (Geng et al., 2019), used a sample-wise level of a small support set to create a general representation of each class utilizing induction networks, this representation is then used in which new queries are compared.

Utilizing pre-trained language models for few-shot learning has also gained increasing attention in recent years (Luo et al., 2021). LEOPARD by (Bansal et al., 2019), applied optimization-based meta learning methods with the BERT model, their framework achieves good performance on diverse text classification tasks. (Luo et al., 2021), implemented meta learning methods to implement BERT with label semantic information for a better generalization performance of few-shot text classification. TARS (Halder et al., 2020), also utilized label semantics to leverage pre-trained language models' performance on different tasks such as sentiment analysis and topic detection for the English language. Their architecture is unified across these tasks with an input of both text with its appending class label and the output is the binary true/false classification based on the matching level between text with its possible label.

3 | MATERIALS AND METHODS

As mentioned earlier, pre-trained language models can efficiently extract discriminative features from the text when utilizing label information. In this section, we demonstrate the effectiveness of using label information on feature extraction from Arabic BERT-based models. Our architecture utilizes fine-tuned AraBERT for the sentiment analysis task and considers the prototypical network as a meta learning framework. The input to this pre-trained language model is modified by appending the label semantic information of ground-truth classes followed by a (SEP) token instead of '(CLS) *tweet* (SEP)' as a common practice of BERT-based classifiers. In other words, the input to the AraBERT model would be as follows:

(CLS) *tweet* (SEP) *label info* (SEP).

The features extracted from the input after augmenting label information, are used to train a linear classifier for text classification. Our method is still a meta learning process for few-shot but the contribution in our work is to analyse whether adding label semantic information will strengthen the capability of AraBERT to extract leveraged representations of short Arabic text.

3.1 | Using BERT models for cross-attention between input text and class labels

As BERT-based models were trained on next sentence prediction (NSP), where predictive information from the next sentence is extracted from the first sentence (Luo et al., 2021). In the case of a text classification task, this could be mimicked when using these pre-trained language models that will extract feature vectors from an input text; by adding label information as the next sentence. The augmented text input will pass through all self-attention layers and in the final layer embeddings are extracted from the (CLS) token. Hence, extracting leveraged features from input text that is relevant to its next class name can be optimized through semantic augmentation of label information. Language models have proven good performance in few-shot text classification tasks without the need of using meta learning methods (Brown et al., 2020), where learning is based on the prior knowledge stored in these pre-trained models (Zheng et al., 2022). However, it has been proven that when appending a class name to an input text, BERT will be able to extract better features from that text (Luo et al., 2021).

3.2 | Proposed model

Our proposed model utilizes the AraBERT encoder to understand the semantics within the text input and the label information. This can be done through the cross-attention mechanism which Transformers' architecture is based on. As described in Section 3, our input sequence is in the order of '[CLS] *tweet* [SEP] *label info* [SEP]'. This sequence is passed through all attention layers of the AraBERT language model. Following the prior work of Luo et al. (2021) and Halder et al. (2020), we used the [CLS] token in the final layer to extract text representations used to train a linear classifier on different text classification tasks including sentiment analysis and stance detection. Figure 4 shows the general architecture of our proposed model.

3.2.1 | Model architecture

Following the work by (Luo et al., 2021), we built a prototypical network meta learner for text classification and this meta learner goal is to find the mapping function between an input text and a class label as described in Section 'Metric learning approach'. However, in our implementation, we augmented this meta learner with label semantics as:

$$\mathrm{f}:(\text{text},\text{label info})\rightarrow\{0,1,2\}^{\mathrm{M}}\ \text{that is},\mathrm{f}(x,d)=\mathrm{P}(ys\mid x,d)\forall \mathrm{s}\in\{1\ldots\mathrm{M}\} \tag{2}$$

To accomplish this goal, a set of support samples is utilized denoted as $X_s = \{x_s, d_s, y_s\}$, where x_s is the input text, d_s is the definition of the ground-truth class label y_s. This set is converted into a classifier $\phi(.; X_s)$ to predict the class label y_q from a set of query samples x_q through $\phi(x_q; X_s)$. Then, to train the meta learner, a meta training data set must be used to randomly construct support sets and query sets considering the *N-way K-shot* setting as $X_s^C = \{x_s^c, d_s^c, y_s^c\}$ and $X_q^C = \{x_q^c, y_q^c\}$, respectively.

The above meta learner is based on a prototypical network (Snell et al., 2017), which is a metric-based few-shot approach, in this approach a limited supervised training set D_{train} consists of N classes each with K examples is used. From D_{train} we randomly generate a support set, then an average of each class feature is used as the class vector w_c for this class. In our work, we can formulate w_c as:

$$\mathrm{w_c}=\frac{1}{|X_s^C|}\sum f(x_s^c,d_s^c), \tag{3}$$

where $(x_s^c, d_s^c) \in X_s^C$ and this feature is calculated after incorporating class label semantic information. Then, on top of the AraBERT encoder, we use a SoftMax function to initiate the probability distribution over the three sentiment classes, that is, positive, negative and Neutral and stance classes, that is, agree, disagree and other.

Then, as described in Section 'Metric learning approach'; each query sample x_q feature vector is mapped to each class prototype w_c using the squared Euclidean distance to find the class with the highest compatibility as the predicted class of the query sample.

The AraBERT pre-trained language model is fine-tuned by the meta learner with added class labels' semantic information to predict the possible label of the query sample y_q, by creating a mapping function in the following form:

$$y_q = argmax\, f(x_q, w_c), \tag{4}$$

where f is the AraBERT feature extractor which converts the augmented input text into an embedding vector. This mapping function is done across each sample in the query set for each class.

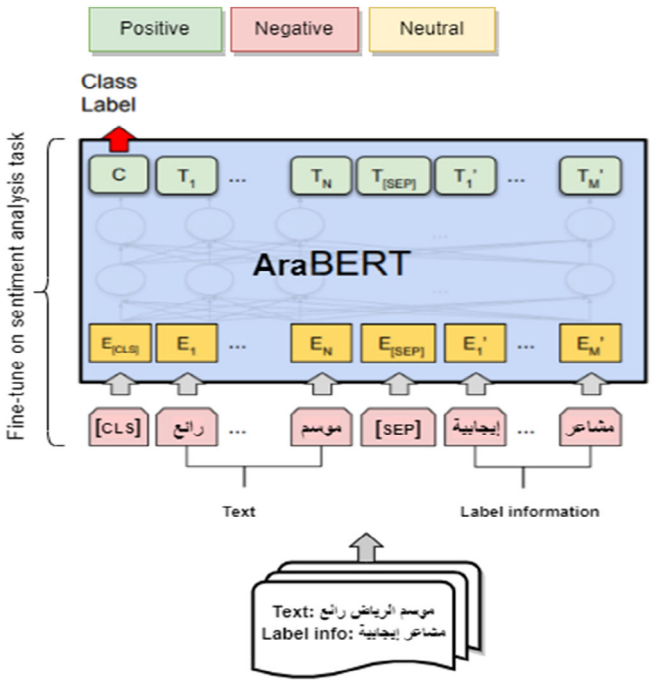

FIGURE 4 Label semantic augmentation model architecture on sentiment analysis task.

4 | EXPERIMENTAL SETUP

In this section, we report our conducted experiments to evaluate the performance of our proposed model which augments short Arabic text with label semantic information in few-shot learning. In our implementation, we employed AraBERT (*bert-base-arabertv02-twitter*) as both the feature extractor and the meta learner. For optimization we utilized the Adam optimizer (Kingma & Ba, 2014), AraBERT learning rate was 2×10^{-5} During training, we split utilized data sets into 80%, 10%, 10% training, validation and testing sets respectively. Then, we employed the episodic training mechanism for non-parametric approaches such as the prototypical networks (Laenen & Bertinetto, 2021), by randomly sampling an increasing number of episodes. In each episode, a support set with N samples per class including 1, 5 or 10 are selected. In addition, Q which is the number of samples of the query set was set to five samples. This selection was based on experiment results, where generally Q is user-defined set in the range of 5 to 15 samples per class (Das & Lee, 2020).

4.1 | Data sets

In our implementation, we aimed to evaluate the effectiveness of augmenting label semantic information on Arabic short-text classification in a few-shot learning. To this end, we experiment with five Arabic data sets: ArSarcasm-v2 (Abu Farha et al., 2021), this data set spans different text classification tasks of sentiment analysis and sarcasm detection and different Arabic dialects with 16% of the text as sarcastic. ArSAS (Elmadany et al., 2018), which is a large data set of 21 K dialectal Arabic tweets from 20 topics in different countries. In this data set, tweets are classified into four-way classifications as either (positive, negative, neutral or mixed) sentiments and speech acts. All tweets were annotated using a crowdsourcing platform with at least three annotators to label each tweet. In the implementation, we followed the work of Abu Farha and Magdy (2021) by ignoring the mixed class, which has the smallest number of samples. Additionally, we ignored labels with low confidence below 50%, this ends up with 18,819 tweets in the data set labelled with three sentiment labels. The third data set is *ArTwitter* (Abdulla et al., 2013), which has two sentiment labels (positive and negative) for Arabic tweets, each label has 1000 examples. However, the online version of the data set has some missing tweets leading to a total of 1000 positive tweets and 975 negative tweets. Another data set used in our experiments was Ans (Khouja, 2020), this corpus consists of contradicted news titles collected from different news resources in the middle east. The data is labelled for the task of stance detection as either (agree, disagree or other). Lastly, *SAtour*, which was self-collected from Twitter about the tourism domain in Saudi Arabia, this data was manually labelled with three sentiments (positive, negative and neutral). Utilizing these data sets will enable experimenting our approach across different domains including sentiment analysis, sarcasm detection and stance detection, where each task has its own labels that hold different semantic information. In addition, we aimed to test our proposed model across different Arabic dialects as provided in these data sets, where the same words can convey different meanings. Table 2 and Table 3 provide some statistics about the online versions of these data sets.

TABLE 2 Data set statistics.

Data set	Task	Class#	Train#	Test#
ArSarcasm-v2 (Abu Farha et al., 2021)	Sentiment, sarcasm	3	12,548	3000
ArSAS (Elmadany et al., 2018)	Sentiment	4	16,415	3482
ArTwitter (Abdulla et al., 2013)	Sentiment	2	1633	342
Ans (Khouja, 2020)	Stance	3	3407	379
SAtour	Sentiment	3	1834	459

TABLE 3 Data set label distribution.

Data set	Positive#	Negative#	Neutral#	Mixed#
ArSarcasm-v2	2577	6298	6495	-
ArSAS	4400	7384	6894	1219
ArTwitter	1000	975	-	-
SAtour	720	731	842	-
Data set	**Agree#**	**Disagree#**		**Other#**
Ans	1301	2399		86

4.2 | Baseline models

We compare our model results against different baselines of support vector machine (SVM), which is an effective model for text classification and regression (Nikam, 2015). And multi-layer perceptron (MLP) which is considered to show high performance and inference speed on long sequence data sets, also this model as a baseline unlike deep learning architectures has a lower operational and maintenance cost (Galke & Scherp, 2021). Table 4 shows the accuracy results of selected data sets using these baseline models.

5 | RESULTS AND DISCUSSION

The results of fine-tuning AraBERT are shown in Table 5, the experiments include feeding AraBERT with different input formats. Since feature extraction in deep learning models is performed implicitly and can result in high accuracy without any explanation (Diwali et al., 2022). Our experiments will include training a classifier with features extracted from Information in the AraBERT decoder versus training a classifier with features extracted from Information provided by label information. Then a final linear layer and an activation function are used to learn on the sentiment analysis task and stance detection task to project a sentiment classification of (positive, negative and neutral) labels or stance classification of (agree, disagree and other). The AraBERT encoder was trained in a few-shot setting of different N shots (1, 5 and 10) for a number of classes $K = 3$ except for ArTwitter with $K = 2$.

From these results, we can conclude that classifiers trained with features extracted from a label information augmented input text, perform better than baseline approaches and from training solely on the input text. This is the case in all selected data sets except for ArSAS, where the performance of an SVM baseline model was .001 slightly better than our proposed model. This might be due to the long sequence input in this data set. For the Ans data set our proposed framework performed better than previous work (Hardalov et al., 2022). Figure 5 shows an observed performance of our approach on ArSarcasm-v2, ArSAS, ArTwitter, Ans and our collected data set SAtour.

TABLE 4 Baseline models accuracy results on utilized data sets.

Data set	SVM TF-IDF	MLP
ArSarcasm-v2	0.639	0.621
ArSAS	0.793	0.733
ArTwitter	0.795	0.793
Ans	0.614	0.562
SAtour	0.620	0.576

TABLE 5 Experiment results of (3-way 1-shot, 3-way 5-shot, 3-way 10-shot) on ArSarcasm-v2, ArSAS, SAtour and Ans data sets, batch_size = 4 except ArTwitter Batch_size = 2 and $K = 2$.

		Number of training samples for each class		
Data set	Label information appended	1 Acc	5 Acc	10 Acc
ArSarcasm-v2	No	0.433 ± 2.0	0.715 ± 2.6	0.751 ± 2.3
	Yes	**0.493** ± 1.7	**0.742** ± 2.2	**0.777** ± 2.2
ArSAS	No	0.645 ± 2.8	0.782 ± 1.5	0.772 ± 1.3
	Yes	**0.658** ± 1.7	**0.792** ± 1.9	0.772 ± 1.1
ArTwitter	No	**0.842** ± 4.4	0.912 ± 4.7	0.915 ± 3.4
	Yes	0.732 ± 2.6	0.912 ± 4.3	**0.954** ± 3.7
Ans	No	0.433 ± 1.8	0.611 ± 1.9	0.873 ± 2.3
	Yes	0.436 ± 1.6	0.667 ± 1.6	**0.886** ± 2.1
SAtour	No	**0.614** ± 4.3	0.792 ± 4.3	0.833 ± 3.9
	Yes	0.592 ± 2.5	**0.800** ± 1.5	**0.863** ± 2.7

Note: As shown in bold values training classifiers with augmenting label information yeilds better results than feeding input text solely.

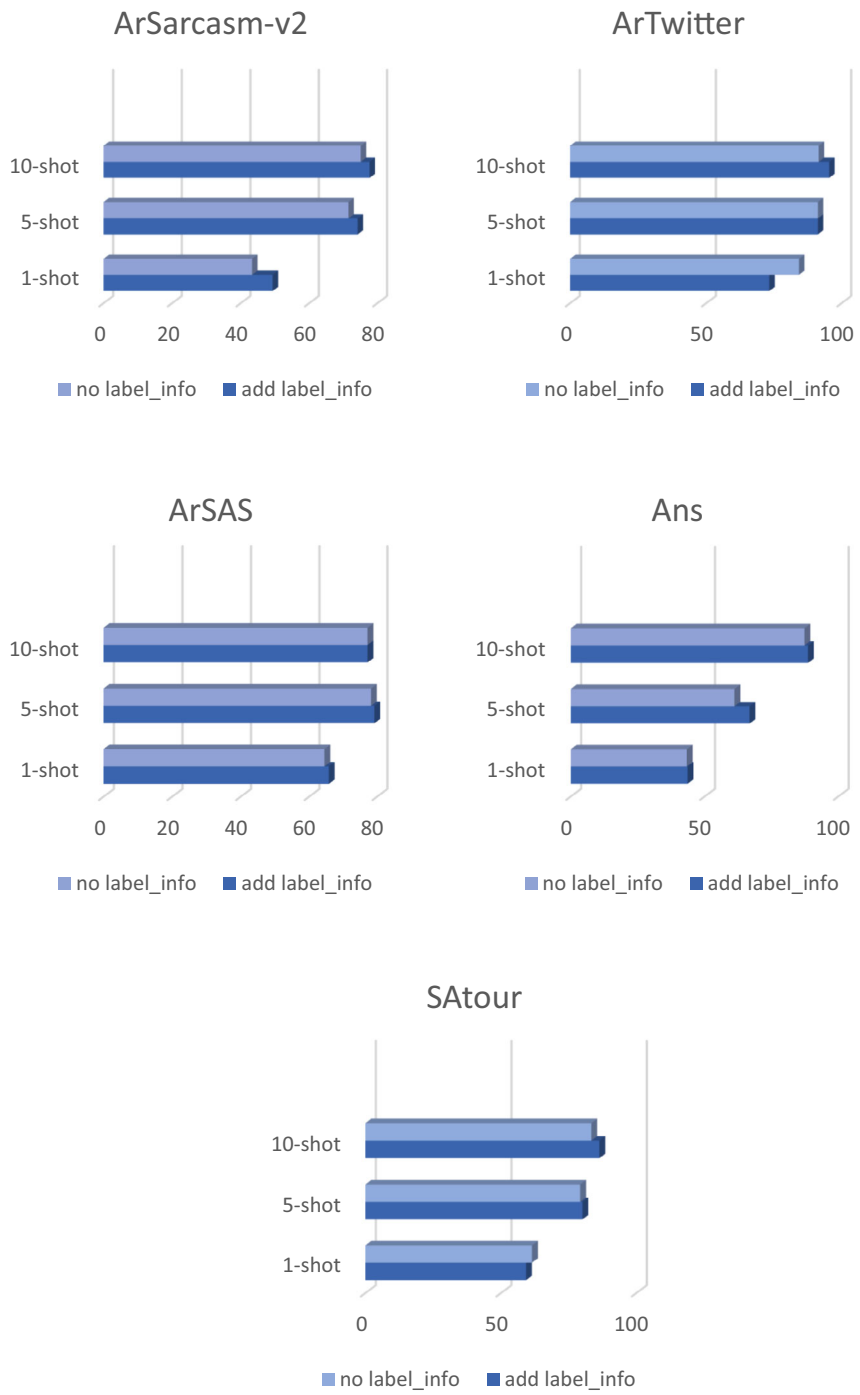

FIGURE 5 Accuracy results on utilized data sets using label semantics versus no label semantics added.

5.1 | Cross-domain zero-shot text classification utilizing learned features

In this section, we experimented short-text classification augmenting label semantics in a zero-shot setting, utilizing the task-aware representation of sentences (TARS) model (Halder et al., 2020). In this approach, the model requires *K* forward passes for each class-input pair, by training the model with a full set of the source task. Then, this model is fine-tuned on the target task using few or no training samples for the target task. This can be achieved by sharing the entire model of encoder and decoder across different tasks which perform the matching between texts and their actual labels of *K* classes. During training, the TARS model should be fed with a number of pairs equal to the number of labels in the resource task for example if the training task is sentiment analysis with two labels the input/output pairs will be as:

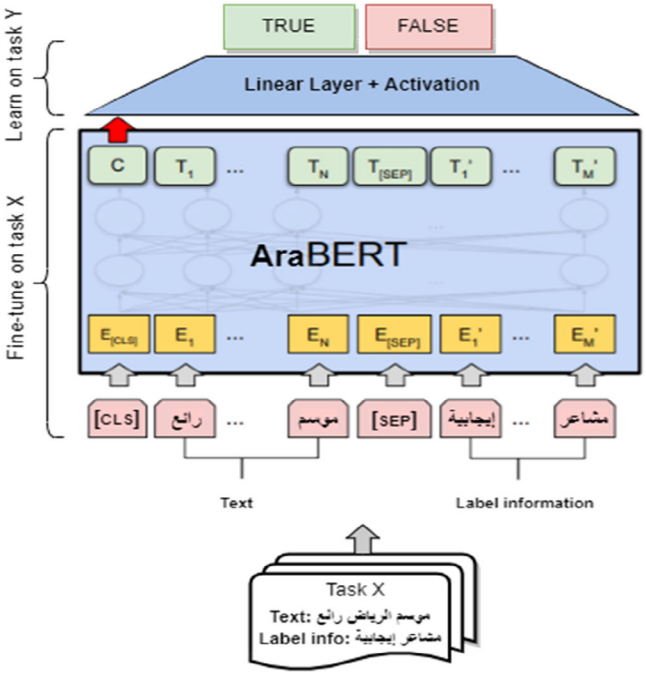

FIGURE 6 Label semantic augmentation model architecture on cross-domain text classification tasks.

TABLE 6 Cross-domain zero-shot text classification experimental results.

Task	Accuracy
ArSarcasm-v2 (sentiment) → ArSarcasm-v2 (sarcasm)	0.874 ± 2.0
ArSarcasm-v2 (sarcasm) → ArSarcasm-v2 (sentiment)	0.452 ± 1.7

< 'I loved the event', 'positive sentiment' > → TRUE.
< 'I loved the event', 'negative sentiment' > → FALSE.

This approach allows zero-shot and cross-domain classification using the TRUE/FALSE decoder instead of the traditional task-specific decoder. In addition, the semantic information provided by labels in resource tasks can be interpreted by the Transformer-based model to leverage the semantics of the new class labels in target tasks. These specifications of matching label semantics with the input text of the TARS model encoder allow sharing of the entire model across different tasks. Figure 6 explains how this approach works.

Using the TARS approach the model was able to classify Arabic short-text across different domains (sentiment analysis, sarcasm detection) with zero-shot training samples of target domain labels. Table 6 shows the experimental results of applying TARS model on classifying Arabic short-text across different classification tasks (sentiment analysis and sarcasm detection) in a zero-shot setting.

6 | CONCLUSION

In this work, we investigate the utility of augmenting label information to Arabic short-text in a few-shot setting with the prototypical network. We conclude that adding the label semantics as an input to an AraBERT encoder can result in more distinct sentence features. Our experimental results demonstrate that we significantly outperform the classification results when training models without augmenting the label semantic information and baseline models. The paper also utilizes the TARS model for zero-shot Arabic text classification across different domains including sarcasm detection and sentiment analysis by matching between the input text and the class label. The model performed well on predicting Arabic text with .874 sarcasm detection accuracy from a model trained on a sentiment analysis task.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their invaluable comments and suggestions. Amir Hussain would like to acknowledge the support by the U.K. Engineering, and Physical Sciences Research Council (EPSRC), his work is under Grant EP/M026981/1, Grant EP/T021063/1 and Grant EP/T024917/1.

CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflict of interest.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY STATEMENT

All data utilized is available online for research purposes.

ORCID

Seham Basabain https://orcid.org/0000-0003-1650-146X
Erik Cambria https://orcid.org/0000-0002-3030-1280

REFERENCES

Abdulla, N. A., Ahmed, N., Shehab, M., & Al-Ayyoub, M. (2013). Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies, AEECT 2013* (p. 6). IEEE. https://doi.org/10.1109/AEECT.2013.6716448

Abu Farha, I., & Magdy, W. (2021). A comparative study of effective approaches for Arabic sentiment analysis. *Information Processing & Management*, *58*(2), 102438. https://doi.org/10.1016/j.ipm.2020.102438

Abu Farha, I., Zaghouani, W., & Magdy, W. (2021). Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the sixth Arabic natural language processing workshop* (pp. 296–305). Association for Computational Linguistics. https://aclanthology.org/2021.wanlp-1.36

Alwaneen, T. H., Azmi, A. M., Aboalsamh, H. A., Cambria, E., & Hussain, A. (2022). Arabic question answering system: A survey. *Artificial Intelligence Review*, *55*(1), 207–253. https://doi.org/10.1007/s10462-021-10031-1

Antoniou, A., Storkey, A., & Edwards, H. (2017). Data augmentation generative adversarial networks. *ArXiv Preprint*. ArXiv:1711.04340.

Bansal, T., Jha, R., & McCallum, A. (2019). Learning to few-shot learn across diverse natural language classification tasks. *ArXiv Preprint*. ArXiv:1911.03863.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chada, R., & Natarajan, P. (2021). FewshotQA: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models. *ArXiv Preprint*. ArXiv:2109.01951.

Chen, Z., Eavani, H., Chen, W., Liu, Y., & Wang, W. Y. (2019). Few-shot NLG with pre-trained language model. *ArXiv Preprint*. ArXiv:1904.09521.

Choi, J., Krishnamurthy, J., Kembhavi, A., & Farhadi, A. (2018). Structured set matching networks for one-shot part labeling. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, *2018*, 3627–3636.

Choudhury, A. (2021). *What are prototypical networks?* Analytics India Magazine. https://analyticsindiamag.com/what-are-prototypical-networks/

Das, D., & Lee, C. S. G. (2020). A two-stage approach to few-shot learning for image recognition. *IEEE Transactions on Image Processing*, *29*, 3336–3350. https://doi.org/10.1109/TIP.2019.2959254

Diwali, A., Dashtipour, K., Saeedi, K., Gogate, M., Cambria, E., & Hussain, A. (2022). Arabic sentiment analysis using dependency-based rules and deep neural networks. *Applied Soft Computing*, *127*, 109377.

Elmadany, A., Mubarak, H., & Magdy, W. (2018). Arsas: An arabic speech-act and sentiment corpus of tweets. *Workshop on Open-Source Arabic Corpora and Processing Tools*, *3*, 20.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, *2017*, 1126–1135.

Galke, L., & Scherp, A. (2021). Forget me not: A gentle reminder to mind the simple multi-layer perceptron baseline for text classification. *ArXiv Preprint*. ArXiv:2109.03777.

Geng, R., Li, B., Li, Y., Zhu, X., Jian, P., & Sun, J. (2019). *Induction Networks for Few-Shot Text Classification* (arXiv:1902.10482). arXiv. http://arxiv.org/abs/1902.10482

Gu, Y., Han, X., Liu, Z., & Huang, M. (2021). Ppt: Pre-trained prompt tuning for few-shot learning. *ArXiv Preprint*. ArXiv:2109.04332.

Halder, K., Akbik, A., Krapac, J., & Vollgraf, R. (2020). Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th international conference on computational linguistics* (pp. 3202–3213). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.285

Hardalov, M., Arora, A., Nakov, P., & Augenstein, I. (2022). Few-shot cross-lingual stance detection with sentiment-based pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 10729–10737.

Hofer, M., Kormilitzin, A., Goldberg, P., & Nevado-Holgado, A. (2018). Few-shot learning for named entity recognition in medical text. *ArXiv Preprint*. ArXiv:1811.05468.

Khalifa, M., Abdul-Mageed, M., & Shaalan, K. (2021). Self-training pre-trained language models for zero-and few-shot multi-dialectal Arabic sequence labeling. *ArXiv Preprint*. ArXiv:2101.04758.

Khouja, J. (2020). Stance prediction and claim verification: An Arabic perspective. *ArXiv Preprint*. ArXiv:2005.10410.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint*. ArXiv:1412.6980.

Laenen, S., & Bertinetto, L. (2021). On episodes, prototypical networks, and few-shot learning. *Advances in Neural Information Processing Systems*, *34*, 24581–24592.

Luo, Q., Liu, L., Lin, Y., & Zhang, W. (2021). Don't miss the labels: Label-semantic augmented meta-learner for few-shot text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP* (Vol. 2021, pp. 2773–2782). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.245

Ma, Y., Cambria, E., & Gao, S. (2016). Label embedding for zero-shot fine-grained named entity typing. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 171–180). The COLING 2016 Organizing Committee.

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental Journal of Computer Science and Technology*, *8*(1), 13–19.

Roy, A., Ghosal, D., Cambria, E., Majumder, N., Mihalcea, R., & Poria, S. (2022). Improving zero-shot learning baselines with commonsense knowledge. *Cognitive Computation*, *14*(6), 2212–2222. https://doi.org/10.1007/s12559-022-10044-0

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, *30*, 4080–4090. https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. *Thirty-First AAAI Conference on Artificial Intelligence*, *31*, 11164.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *2018*, 1199–1208. https://doi.org/10.1109/CVPR.2018.00131

Tang, Z., Wang, P., & Wang, J. (2020). ConvProtoNet: Deep prototype induction towards better class representation for few-shot malware classification. *Applied Sciences*, *10*(8), 2847.

Vanschoren, J. (2018). Meta-learning: A survey. *ArXiv Preprint*. ArXiv:1810.03548.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, *29*, 2016. https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html

Wang, S. (2022). *CS583: Deep Learning* [TeX]. https://github.com/wangshusen/DeepLearning/blob/c36b39215b5adf0694efa6f61f15f07300a725ea/Slides/16_Meta_1.pdf

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2021). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, *53*(3), 1–34. https://doi.org/10.1145/3386252

Wu, S., Zhao, X., Yu, T., Zhang, R., Shen, C., Liu, H., Li, F., Zhu, H., Luo, J., & Xu, L. (2021). Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. *ArXiv Preprint*. ArXiv:2110.04725.

Xia, M., Artetxe, M., Du, J., Chen, D., & Stoyanov, V. (2022). Prompting ELECTRA: Few-shot learning with discriminative pre-trained models. *ArXiv Preprint*. ArXiv:2205.15223.

Xu, Z., Wang, C., Li, P., Li, Y., Wang, M., Hou, B., Qiu, M., Tang, C., & Huang, J. (2021). When few-shot learning meets large-scale knowledge-enhanced pre-training: Alibaba at FewCLUE. In *CCF international conference on natural language processing and Chinese computing* (pp. 422–433). Springer-Verlag.

Xu, Z., Wang, C., Qiu, M., Luo, F., Xu, R., Huang, S., & Huang, J. (2022). Making pre-trained language models end-to-end few-shot learners with contrastive prompt tuning. *ArXiv Preprint*. ArXiv:2204.00166.

Yu, M., Guo, X., Yi, J., Chang, S., Potdar, S., Cheng, Y., Tesauro, G., Wang, H., & Zhou, B. (2018). Diverse few-shot text classification with multiple metrics. *ArXiv Preprint*. ArXiv:1805.07513.

Zheng, Y., Zhou, J., Qian, Y., Ding, M., Liao, C., Li, J., Salakhutdinov, R., Tang, J., Ruder, S., & Yang, Z. (2022). *FewNLU: Benchmarking State-of-the-Art Methods for Few-Shot Natural Language Understanding* (arXiv:2109.12742). arXiv. http://arxiv.org/abs/2109.12742

AUTHOR BIOGRAPHIES

Seham Basabain received the B.Sc. degree in 2012 from the department of Information Systems, Faculty of Computing and Information Technology at King Abdul Aziz University, Jeddah, Saudi Arabia. And the M.Sc. degree in 2015 with distinction from the School of Computing, Faculty of Engineering at the University of Leeds, Leeds, United Kingdom. She is a lecturer in the department of Information Systems at King Abdul Aziz University and currently pursuing the Ph.D. degree from the school of computing, Edinburgh Napier University, Edinburgh, United Kingdom. Her current research areas include natural language processing, Arabic text analysis, semantic information, feature analysis, few/zero-shot learning and semi-supervised learning.

Erik Cambria received the Ph.D. degree in computing science and mathematics from the University of Stirling, Stirling, Scotland, in 2012, following the completion of an EPSRC project in collaboration with MIT Media Lab., He is currently an Associate Professor with Nanyang Technological University, Singapore, where he also holds the appointment of Provost Chair in Computer Science and Engineering. He is the Founder of Sentic Net, a Singapore-based company offering B2B sentiment analysis services. His research focuses on neuro symbolic AI for explainable natural language processing in domains, such as sentiment analysis, dialogue systems, and financial forecasting. Dr. Cambria was the recipient of many awards, e.g., the 2018 AI's 10 to Watch and the 2019 IEEE Outstanding Early Career Award, and was featured in Forbes as one of the 5 People Building Our AI Future.

Khalid Alomar received his B.Sc. in Software Engineering from University of Brighton, U.K., in 2003. Then, he received his M. Sc (with distinction) and Ph. D degrees in Software Engineering, from the University of Bradford, U.K., in 2005 and 2010, respectively. Currently, he is an Associate Professor in the Department of Information Systems at King Abdul Aziz University in Saudi Arabia, where he served in several leadership positions, including Vice Dean for Technical Affairs at the Deanship of e-Learning and Distance Education for 10 years. His research interests are human-computer interaction and machine learning.

Amir Hussain received the B.Eng. and Ph.D. degrees in electronic and electrical engineering from the University of Strathclyde, Glasgow, U.K., in 1992 and 1997, respectively. Following post-doctoral and senior academic positions at the University of the West of Scotland, Paisley, U.K., from 1996 to 1998, the University of Dundee, Dundee, U.K., from 1998 to 2000, and the University of Stirling, Stirling, U.K., from 2000 to 2018, respectively, he joined Edinburgh Napier University, Edinburgh, U.K., as the Founding Head of the Cognitive Big Data and Cybersecurity (CogBiD) Research Laboratory and the Centre for AI and Data Science. His research interests include cognitive computation, machine learning, and computer vision.

How to cite this article: Basabain, S., Cambria, E., Alomar, K., & Hussain, A. (2023). Enhancing Arabic-text feature extraction utilizing label-semantic augmentation in few/zero-shot learning. *Expert Systems*, e13329. https://doi.org/10.1111/exsy.13329