

RESEARCH ARTICLE

Feature Selection for Location Metonymy Using Augmented Bag-of-Words

MUHAMMAD ELYAS MEGUELLATI¹, ROHANA BINTI MAHMUD¹,
SAMEEM BINTI ABDUL KAREEM¹, ASSAAD OUSSAMA ZEGHINA², AND YOUNES SAADI¹

¹Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

²Icube Laboratory, Université de Strasbourg, CNRS, ICube UMR 7357, F-67000 Strasbourg, France

Corresponding author: Muhammad Elyas Meguellati (meg.moh.ilyes@gmail.com)

This work was supported in part by Universiti Malaya Research University Grant-Faculty Program under Grant GPF091A-2020, and in part by the Trans-Discipline Research Grant Scheme Program TR001D-2018A.

ABSTRACT Location metonymy resolution is a study that deals with locations being used in a non-literal way that create problems in several natural language processing tasks such as Named entity recognition and Geographical parsing. Many studies were conducted attempting to accurately classify whether the location is used literally or metonymically, however, most of the approaches that performed well had to employ a considerable amount of resources along with complex machine learning models; those that reduced the resources experienced a decline in performance due to data sparseness. This study proposes a novel feature selection approach that uses bag-of-words and augments it with GloVe embeddings to obtain features that can be recognized based on the context of the sentence. We then implement a minimalist deep learning model making the entire classification task as light as possible. The study found that relying solely on the given datasets to identify features without depending on other external resources can achieve remarkable results despite the small size of the datasets. The results obtained from evaluating our method compared to the state-of-the-art methods show that eliminating noise based on the context notwithstanding the usage of low-cost resources has outperformed all of the previous methods with an accuracy of 99.2% on the WIMCOR dataset.


INDEX TERMS Text classification, metonymy resolution, deep learning, feature selection, bag-of-words, natural language understanding.

I. INTRODUCTION

Our everyday language is filled with expressions that are used in a non-literal way which we humans can easily interpret; however, some figurative speech can be a challenging task for a machine to interpret especially where there are no visible insights to rely on. Metonymy is one of the figures of speech that do not have a surface structure, it stands for the use of a word or an expression to refer to a concept with a close association [1]

Example 1: “Here comes the gun.”

In Example 1, the word “gun” is not used to refer to the object that shoots bullets, but to the person who is known to be a good shooter with the object “gun”.

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

Our primary objective in this paper is to address a type of metonymy that deals with location names. Location Metonymy is a subcategory that expresses using a location name for anything other than a geographical place.

Example 2: “Germany won the world cup in 2014”.

As shown in example 2, the word “Germany” does not refer to the geographical location where the country Germany is situated, however, it refers to the team of *players* that share the German nationality, thus, its actual reference is linked not to the place but to the people.

Several Natural Language Processing (NLP) tasks do not have the ability to distinguish between a location being used literally or metonymically such as Named Entity Recognition (NER), Geographical Parsing (GP) and Machine Translation (MT) [2], [3]. According to Gritta *et al.* [3], location metonymy covers one fifth of the locations mentioned in

Wikipedia, making information retrieval (IR) applications prone to having a faulty observation 20% of the time when it comes to identifying the sense in which a location is used in, therefore, accurately classifying the location mentioned to either literal or metonymy can significantly decrease the margin of error in such NLP applications. Two of the major obstacles that occur when viewing metonymy as a classification task are: Firstly: the generalization conundrum, where it is significantly complicated to categorize together the types of metonymy that follow approximately the same pattern (as in the case of locations and organizations) due to the endless possibilities of senses a word can take. Secondly, data sparseness: datasets for this task are relatively small compared to other NLP tasks, which causes a limitation to some extent during the feature extraction phase. Taking into account these limitations, researchers (particularly the participants of SemEval 2007 Shared Task 8) tend to exploit external resources such as WordNet, FrameNet, Multiple Parsers...to assist with constructing a heavily engineered features alongside a sophisticated machine learning model making the whole approach costly.

The primary objective of this study is to demonstrate that it is possible to reduce the resources employed in previous research without compromising the performance of the classification task relying only on the dataset, GloVe word embeddings and a minimalist deep learning model with a convolutional neural networks (CNNs) architecture.

II. RELATED WORK

A. DATASETS AND EVALUATION

One of the well-known issues that mainly occurs when dealing with tasks that can be quantified is *data invariance* [4]. Data invariance is assessing how the results can change when changing the dataset that is applied to a specific method; Metonymy Resolution is one of the NLP tasks that had the issue of *non-standardized* dataset which stood in the way of progressing, as a consequence, several studies have tried to either construct a customized sample with a few sentences [5], [6] or to use different corpora [7], which makes the results obtained by the different studies hardly comparable, thus, difficult to *objectively* follow the advancement of the *state-of-the-art*.

B. SMALL-SCALE METONYMY SYSTEMS

Before the SemEval dataset was introduced in 2007, Metonymy resolution early work was largely based on violating the *selectional preferences* [8]–[10]; yet by focusing on selectional preferences only, a considerable number of metonymic readings were skipped [11]. On the other hand, *Schematic Metonymy* can be easily generalized to be a corpus-based work such as location metonymy and organization metonymy, allowing it to be suitable for using statistical approaches along with semantic understanding.

Small-scale metonymy systems primarily focus on *Logical Metonymy* which does not necessarily comply with *Schematic*

Metonymy, however, investigating the earlier approaches can yield a better understanding. One of the notable studies in Logical Metonymy was conducted by [7], which highlighted the shortcomings of describing all the possible interpretations to logical metonymy and how the ranking of interpretations is not based on the likelihood of the metonymy, for those reasons [7] identified the sense of the metonymic words from a large corpus (British National Corpus) by extracting the collocation and co-occurrence then applying it on a probabilistic model to rank the interpretations accordingly.

Shutova and Teufel [6], approached the issue of logical metonymy by presenting a statistical method, gathering collocation “synsets” and clustering verbs based on similarities, these sets of features help logical metonymy detection, however, one of the conceivable flaws [4], is that neither scale invariance nor dataset invariance was mentioned, their K-means clustering method was trained on five sentences and tested on five others, additionally, the sentences were created by the authors.

C. LARGE-SCALE METONYMY SYSTEMS

Before approaching Metonymy as a classification task, researchers categorized it into two main classes “unconventional” and “schematic”. The key difference between both classes [12] is the ability of the schematic to be at some point generalized, that means it can be used in the same way in many different contexts for example, “Africa speaks”, it indicates that the location name “Africa” is used to represent African people. However, the unconventional metonymy is more creative, therefore, it is likely to be used for a unique situation and hardly repeated which makes it difficult to be recognized (e.g., “Finish your plate”), the literal interpretation of this example is “Finish eating all the food in your plate”, as humans, it is evident that the demand is to finish eating the food, however, the way the machine interprets it will be word-for-word, thus, the sense will be something like: to kill the plate or break the plate. The primary focus of the following review is “The schematic metonymy” more precisely the location names metonymy where geographical names are used in a non-literal way.

1) SemEval 2007 PARTICIPANTS

Nissim and Markert [13], study the syntactic features and word similarities for metonymy resolution with a supervised approach, the syntactic head-modifier relations and its correlation with precision, a solution to overcome the sparseness of data by presenting simplified grammatical features and generalize over two levels of contextual similarities. They introduce an annotated corpus of occurrences of country names by differing between the metonymic patterns that follow: (place-for-people) a location that stands for an organization or a person associated with it e.g. “Scotland lost the semi-final”, (place-for-event) a place that refers to an event that has been occurred there e.g. “The Afghanistan was very tough” where Afghanistan is referring to the event “war” and not the actual location, (place-for-product) a location where a specific and

famous product is manufactured in it e.g. “The Bordeaux never changes” where Bordeaux is referred to the wine produced locally in Bordeaux, there is also another category that covers the other metonymic expressions that failed to fall under one of the three previous patterns called “othermet” [14], and finally, the Mixed category where the location name can be interpreted both literal and metonymy at the same time such as: “U.S army arrived to Iraq, they wanted to force their way into the USA embassy yesterday” in this example “Iraq” can be both a place and people at the same time. However, due to the rareness of the othermet category they reformulate the classification task between literal and metonymy only by using Decision List (DL) classifier where the whole features that were encountered during training are ranked in the DL, they used Backing-off strategy for the literal locations mentioned where no decision can be made. Their study shows that decreasing the context of head-modifier relations achieves high precision in recognizing metonymies (74.5%) and (90.2%) accuracy, which indicates that the process of reducing does not cause the loss of critical information, and they also show generalizing context similarity increases the recall without any increase in the training set size nor sacrificing precision.

More of metonymy resolution comes from the SemEval 2007 Shared Task 8 and later by Nissim and Markert [15], where the features set was updated from Nissim and Markert [12] work such as the lemmatized head-modifier and the grammatical roles features. Due to the broad level that metonymy has, Nissim and Markert [16], carry on their previous study on recognizing metonymy by exploring other patterns of it and annotating it as well, the explored patterns fall under the organization category. The standard sense of the word “organization” describes it as a reference to the organization in general which by definition is a legal entity that contains organization members who speak for the organization and has a defined goal (e.g., “Intel’s new CPU is powerful, yet not too expensive”). On the other hand, the metonymic sense consists of six types: 1 org-for-members, 2 org-for-event, 3 org-for-product, 4 org-for-facility, 5 org-for-index, 6 othermet and also a mixed category. In order to annotate the data, they use CIA Factbook as well as the fortune of five hundred list as samples of countries and companies, then they pass the annotated data to the teams participating in the SemEval 2007 to classify metonymies at three different levels (coarse, medium, fine), the teams use supervised and unsupervised approaches to classify the given data and the performance is as follow: in “location- coarse” the accuracy baseline is 0.794 and the average is 0.815, however, “organization-coarse” category scored less with a baseline accuracy of 0.618 and an average of 0.718. Farkas *et al.* [17], use Maximum Entropy strategy in order to enable the model to learn, setting the Gaussian prior to 1 and using random five-fold cross validation to define the importance of a specific feature because of the limited amount of features, therefore, the learning algorithm converges rapidly without the need of too many iterations

(less than 30). They also use the C4.5 decision tree algorithm but when comparing, the Maximum Entropy model always performs better. Ferkas *et al.* [17], make use of feature engineering (1-grammatical annotations, 2- determiners, 3-grammatical numbers) provided by Nissim and Markert [12], Their result is the leanest among the teams participating in the SemEval 2007 Task with an accuracy of 85.2%.

During the SemEval 2007 Task 8, Brun *et al.* [18], is the only team that uses an unsupervised approach for the Named Entity Metonymy Resolution by combining a robust deep parser and lexical syntactic information, with a distributional method using the similarities of the syntactic content calculated on a large corpus. Their system uses the Xerox Incremental Parser (XIP) to perform a deep syntactic analysis by constructing a set of syntactic relations from a text, these relations are labeled with deep syntactic predicate (verbal and nominal) with deep Sub, deep Obj and modifiers. In order to adapt their system to the SemEval 2007 Task 8, they adjust the parser so it would be able to handle Named Entity Metonymy and by following the guidelines of the SemEval 2007 Task 8 set by [16], they perform a corpus study to extract, lexical and syntactical features from the data for both location and organization, for instance, the organization mentioned in the input text is the subject of the sentence and the following verb is an economy-related action such as pay, refund, export, etc. . . , it is highly likely to be org-for-members. Their results are significantly higher than the baseline for both location and organization names, they score 85.1% accuracy compared to the baseline 79.4%.

2) RESEARCH AFTER THE SemEval

On the other hand, Nastase and Strube [19], try to free the metonymy resolution task from the complication of data sparseness by going beyond the familiar approaches, they use unsupervised learning feature to their supervised approach to allow them to go beyond annotated data, applying techniques which are generally used for Word Sense Disambiguation (WSD) unsupervised learning in order to pass restrictions of the grammatical relations neighbors by collocations extracted from corpus, therefore, the single annotation needed is the Part Of Speech tags that enables them to retrieve grammatical relations using the British National Corpus (BNC) to define the type of subject/object preferred in “Word-POS: frequency” format, then determining the interpretation accordingly e.g. person, product, facility. . . They use WordNet 3.0 to pick the closest possible reading. As a result of their method they achieve 86.1% accuracy.

Nastase *et al.* [20], expand the approaches used when the constraints are placed in a Possible Metonymic Word (PMW) by its grammatical collocates taking into account the larger context, along with the local grammatical context, however, a map is constructed as a concept network concepts by transforming Wikipedia into a large-scale multilingual concept called WikiNet, which gives access to the conceptual relations of the PMW and the other concepts, for the local context metonymy resolution they use selectional preferences

derived from semantic relations from the concept network and what surrounds it globally as an input to be interpreted, their approach that uses a probabilistic framework for the local and the global context and extracting concepts from WikiNet significantly beats the baseline, this indicates that the relations extracted are practical substitute from the classes provided manually for an unsupervised framework.

Nonetheless, recently, Gritta *et al.* [3], approach the problem from a different angle, they view that the SemEval 2007 dataset has unbalanced classes (79.7% literal, 19.3% metonymy, 1% othermet) which signifies that the bias is too high for the classification task, therefore, they create a new dataset “ReLocaR” with the help of linguists from Cambridge university that has (49% literal, 49% metonymy). Besides their newly created dataset, they develop a new approach to extract features that are around the PMW, unlike the immediate5 and immediate10, their approach uses a predicate window of 5 words “PreWin” as they believe that most of the sentence is irrelevant to the PMW, therefore, the extraction of unrelated words would have a negative effect on the trained model, after extracting the features using their novel method, they apply a deep learning approach using LSTM (long short term memory). The results show that their method works well on the three datasets they use (SemEval 2007, ReLocaR, CoNLL) given that they implement a minimalist model compared to precedent approaches, where they achieve 83.1% accuracy when using the SemEval 2007 dataset and 83.6% accuracy when using ReLocaR.

Owing to either the insufficient coverage or the sparseness of SemEval and ReLocaR datasets, Mathews and Strube [21], present a new dataset for location names called WIMCOR (Wikipedia Metonymy Corpus), the samples were generated semi automatically using Wikipedia Disambiguation Pages a (Auer *et al.*, 2007; Lehmann *et al.*, 2015) which is a list of the different senses a one word can be used in, and offers links to articles that hold each meaning. One of the two main procedures to generate the dataset is by using metonymic pairs where the word is the same with two different concepts that are strongly connected, the intuition behind how to determine whether the concepts are highly linked is: “the more links referring to one another, the stronger the connection between the pair”. Furthermore, they use Wikipedia to extract samples and apply some constraints such as restricting the length of the sample in the range of 10 to 512 tokens. After the generation of WIMCOR dataset, the authors test the baseline models on it, however, their metrics of measurements are different from most of the studies, they use precision, recall, f1-score, whereas, accuracy is the primary evaluation metric in earlier research.

Given that, most of the previous studies were released before 2019, BERT (Bidirectional Encoder Representations from Transformers) had not been released yet. Li *et al.* [22], replicate Gritta *et al.* [3], approach using BERT which achieved substantially higher results. PreWin using GloVe achieved 83.1% and 83.6% for SemEval and RelocaR respectively, PreWin using BERT scored 87.1% and 92.2%.

Additionally, they use BERT as their model for the classification task with three different settings BERT base, BERT LG (large) and BERT MASK (they mask the PMW from the training). The results of their experiments achieve the state of the art with 89.1%, 94.8% and 95.9% in SemEval, ReLocaR and WIMCOR respectively. Du and Wang [41] state that the *state-of-the-art* methods only focus on the context of the sentence neglecting *Entity representation* and *Syntactic structure*, their study is fixated on exploiting entity and syntax constraints by obtaining syntactic dependency relations, then developing a neural network that can integrate both constraints for a better representation, thus, enabling the model’s performance to surpass the limitations when encountering complex sentences along with reducing the noise. The experiments on SemEval and ReLocaR datasets display a considerable improvement where it scored 89.8% and 95.7% respectively which is over 4% compared to the BERT model [22].

TEXT CLASSIFICATION WITH CNN-BASED MODELS

Many NLP tasks lean towards implementing Recursive Neural Networks (RNNs) architecture as their deep learning model e.g. question-answering, machine translation, part of speech tagging, etc...due to its ability to view text as a sequence which allows it to capture patterns over time, in contrast, CNN are capable of identifying patterns throughout space as LeCun *et al.* [23], put it: RNNs are accustomed to recognize patterns over time, CNN are accustomed to recognize patterns over space. Given that RNNs process text as a sequence of tokens to locate dependencies and form structures which are often unnecessary in text classification (TC) tasks, ergo, CNN have become one of the widely used architectures in TC for the reason that they are capable of spotting local and global context. Kalchbrenner *et al.* [24], is among the earliest who relied on CNN for capturing relations and predicting sentiments with their Dynamic CNN

(DCNN), additionally, it can be employable to all languages. The performance of their networks achieve high results on Twitter sentiments without the need of external resources (25% error reduction). Alternatively to the previous study, Kim [25] CNN model is less complex than DCNN where their approach uses a 1D convolution filter on the word vector that was created by one of the word embeddings algorithms such as GloVe. Kim [25], furthermore evaluates several approaches of word embeddings (CNN-Rand, CNN-static, CNN-non-static, CNN-multi-channel), these methods recorded to improve the state of the art in four tasks out of seven which establish that pre-training of word vectors prior to a deep learning model can significantly increase the performance. Liu *et al.* [26], approached the problem of “Extreme Multi-Label Text Classification” (EMTC) as the first attempt that uses deep learning model, the CNN architecture that was proposed yielded notable results where it accomplished either the best or the second best across all datasets of the study. Johnson and Zhang [27] experiment applying CNN on high

dimension text to leverage the 1D structure which led to learn embeddings of small text regions.

Driven by the work of Simonyan and Zisserman [28] and He et al. [29], Conneau et al. [30], presented a “Very Deep – CNN” (VD-CNN) which perform convolutions and poolings on the character level, the study demonstrates the increase in performance in relation with the depth if the model. According to Duque et al. [31], the primary objective of most research in CNNs is to boost the accuracy by incrementing the networks’ depth which resulted in creating a numerous amount of parameters that negatively impacted the memory and processing during training. Their proposed approach “Squeezed Very Deep CNN” (SVD- CNN) is designed to fit mobile platforms, it is 10 to 20 times smaller with a minor decrease in accuracy (0.4% - 1.3%). Le et al. [32], investigated the impact of Deep CNN in text classification, their research shows that on the character level deep CNN out perform shallow and wide CNN, contrarily, they found that shallow and wide CNNs outperform deep CNNs when the sequence input is a word. Zhang et al. [33], conducted a sensitivity analysis on the input passed to a CNN layer, their analysis found that one-hot vector performs poorly for a moderate to small datasets size unlike word2vec and GloVe embeddings.

Many other remarkable text classification with a CNNs-based model are not covered in this section, nonetheless, those mentioned have a direct impact on the work that is done later on in the methodology and the experiments.

III. APPROACH

This section highlights the datasets used in the experiments alongside the different phases our approach consists of as shown in Fig. 1.

1- Starting with a preprocessing phase to the selected dataset in order to eliminate certain stopwords. 2- We then identify the most repeated words in each category to form the preliminary bag-of-words. 3- Using GloVe, we augment the previous BOW to contain more similar words to the initial ones. 4- We preserve the intersection of the Augmented BOW with the samples of the dataset. 5- We pass on the input to our deep learning model for training. Finally, we evaluate the results obtained using the suitable metrics of measurement.

A. DATASETS

1) SemEval 2007

One of the obstacles that stand in the way of advancement in Location Metonymy Resolution is data sparseness, it was until 2003 when the first dataset was created by Markert and Nissim [14] then introduced by the same authors in the Semantic Evaluation (SemEval) 2007 conference as “SemeEval 2007 dataset”. The process executed for the dataset creation starts by extracting all country names from WordNet [34] and the CIA fact book (<http://www.cia.gov/cia/publications/factbook/>) to form a list called “CountryList”, following that, from the BNC (The British National Cor-

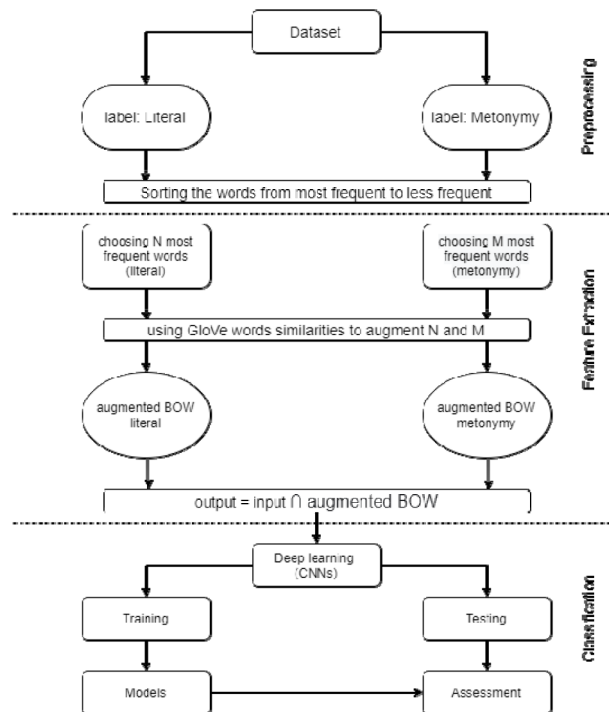


FIGURE 1. Steps of the proposed framework architecture, where it shows the different phases described in this paper.



FIGURE 2. The original two words from each category in the initial BOWs alongside the closest 4 words represented by GloVe.

pus info.ox.ac.uk/bnc) 1000 samples that include any of the countries from CountryList are taken, few of the samples are filtered out then the manual annotation begins, The PMW is categorized to either literal or metonymy, additionally, the authors divided the metonymic class into four subclasses (1- place-for-people, 2- place-for-product, 3- place-for-event 4- othermet). The reproducibility of the results has a percentage agreement of 0.95, and a Kappa of 0.88 [35], [36], signifying that the corpus can be considered reliable. After eliminating the samples on which both computational linguists did not agree, the final number of samples (training and testing) is currently 1833 samples.

2) ReLocaR

Due to the shortcomings in SemEval 2007 dataset particularly the imbalanced classes and the different annotation scheme (where the political entities are considered as literal

locations) according to Gritta *et al.* [3], ReLocar was introduced with some improvements compared to the previous dataset. The main differences that ReLocar comes with are: 1- the samples were extracted from Wikipedia, 2- both classes occupy 49% of the total number of samples, 3- the political entity is considered as a metonymic reading. Similar to SemEval dataset, ReLocar was manually annotated by four trained linguists and two computational linguists from The University of Cambridge.

3) CoNLL

It was presented alongside ReLocar [3], Samples (locations only) from the CoNLL 2003 NER Shared Task data were annotated with a lack of quality labels, therefore, the dataset is more susceptible to noise. CoNLL dataset has a total of 7057 samples divided as 4609 literal and 2448 metonymy. Compared to SemEval 2007 and ReLocar datasets, it is almost twice the size which can be advantageous during training.

4) WIMCOR

Unlike the earlier datasets, Mathews and Strube [21], attempted to semi automate the process of generating a dataset for location metonymy resolution, using “Wikipedia Disambiguation Pages” the authors extracted the different meanings a word can have then categorizing each word with its corresponding article from DBpedia [37], [38]. The dataset contains more than 1000 unique PMW, however, in contrast to the number of samples (206,000 samples) it is considered small, the average length of a sample is 80 tokens. The statistics of WIMCOR and the previously mentioned datasets can be observed in Table 1.

TABLE 1. The datasets statistics used in the study. AVG is the average number of words in a sample.

Dataset	literal	Metonymy	PMW	AVG	Source
SemEval	1458	375	262	26.6	BNC
ReLocar	955	1031	603	22.7	Wikipedia
CoNLL	4609	2448	1685	24.6	CoNLL
WIMCOR	154322	51678	1029	85.3	Wikipedia

B. DATA PREPROCESSING

We begin by manually selecting the stop words in order to prevent the NLP pre-existing library of stop words to remove the words that may indicate the category of the PMW, given that, we are dealing with locations, prepositions such as “from, to, above, below...” need to be preserved. After removing the unnecessary words, we create two lists, a list that contains all the samples that are labeled literal and a list with all the samples labeled metonymy. We then extract from each list the unique words and their number of occurrences. Observing the words with the high frequency in each class, we form two preliminary bag-of-words: metonymy and lit-

eral, each bag holds some of the most repeated words in that same category.

C. BAG-OF-WORDS AUGMENTATION

Now that both bag-of-words are established, considering that GloVe word embeddings algorithm Pennington *et al.* [39], group together the words’ vectors that are similar or co-occur frequently as shown in Fig. 2, we use it to obtain the nearest words’ vectors to each word in BOW (as shown in figure 1) excluding those that do not require augmentation e.g., for, by, with, in, etc). This added step is to include the words that may have the same significance as the ones obtained earlier (i.e. the most frequent words) but are not used too often in the dataset, in order to widen the range of context words for each category.

D. FEATURE SELECTION

With the help of GloVe word embeddings algorithm, the augmented bag-of-words (A-BOW) have over 100 words for each category. Our feature selection approach is mainly designated to eliminate maximum noise regardless to whether the sentence’s meaning is preserved or not, as the main objective is to determine the nature in which the location is used in. Note that as a result of using GloVe for augmentation some words may not exist in the dataset, however, the feature selection starts by finding the intersection between the A-BOW and the dataset samples as shown in Table 2, this process eliminates the possibility of having words out of the dataset’s vocabulary. Prior to passing on the end results, we verify that the length of the intersection is at least two words (theoretically, one word can be used in any sense, therefore, it cannot identify the context), the samples that did not fulfill the condition are replaced by the immediate 10 approach which means taking 10 words from the left and 10 words from the right of the PMW.

TABLE 2. The intersection output between the original samples with A-BOWs for each category on ReLocar dataset.

Literal samples	Metonymic samples
to eastern	led by
population in on the of	for to diplomatic
in from to empire	during soviet union diplomatic to
in cities of	of its neighbors and Ceylon
on be to months	independence for
city in central	to diplomatic
well of most	Europe with
in of provinces	match tournament against
population of in city state	both formally recognized on that
south	same
since in states from	played for during playing

E. DEEP LEARNING MODEL

Deep learning models have outperformed the common machine learning approaches in several tasks according to Minaee *et al.* [40], who have made a comprehensive review for over 150 deep learning-based models regarding various

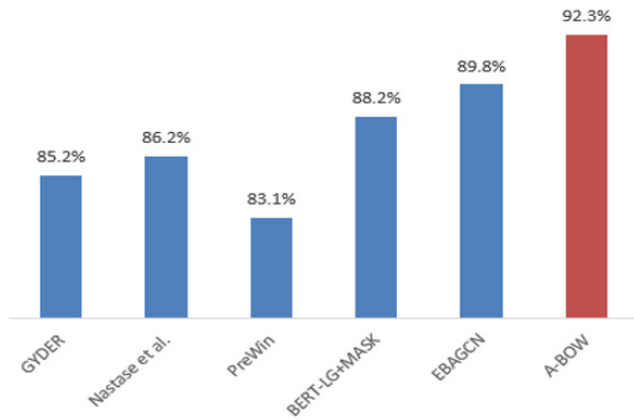


FIGURE 3. The accuracy of some of the previous approaches compared to A-BOW on the SemEval dataset.

text classification tasks e.g. sentiment analysis, news categorization, question answering.

The output of the feature selection phase is used to train the deep learning model. The reason we chose CNNs architecture is for how fast the training is compared to other models such as RNNs. Contrary to RNNs, CNNs lose the local order of information about the words due to the operations of convolution and pooling making it unfit for tasks that require the order of the words to be preserved such as dependency parsing and part-of-speech tagging, however, the information about the order is not crucially necessary in many text classification tasks like Location metonymy classification. Additionally, with a large vocabulary, the cost of computing becomes exponentially expensive specifically with models that apply recursion (i.e. LSTM and Gated Recurrent Unit (GRU)), whereas, Convolutional filters have the ability to automatically recognize representations without the full vocabulary.

The output of A-BOW is used to train the following model. The input with a sequence length of 24 is passed to the embedding layer using GloVe as the weight, followed by a 1D convolutional layer of 64 filters and 12 as the filter's size, then, a max pooling layer with a pool size of two. Three fully connected (dense) layers of 32,16,1 respectively. Relu function is used in all the layers as the activation function except the last layer in which we used a Sigmoid activation. A Binary Cross Entropy is employed to train the model as a post function with a learning rate of 0.0001. It is important to mention that the results obtained during training are highly similar to the test results we will be discussing in the results section.

IV. RESULTS

Following the SemEval 2007 convention, the metrics that are used for the experimental evaluation are accuracy and f1-score. The evaluation of the performance is at the *coarse level* where the two classes of *metonymy* and *mixed* are combined into one class "*non-literal*". In this section, we divide the discussion of the results achieved into two parts, the first part will highlight the performance of our feature selection

method compared to the baseline methods, where both our study and the baseline will be trained on the same model to neutralize the effect of having different training models. The second part is a discussion of the entire location metonymy model's performance in comparison with the previous models.

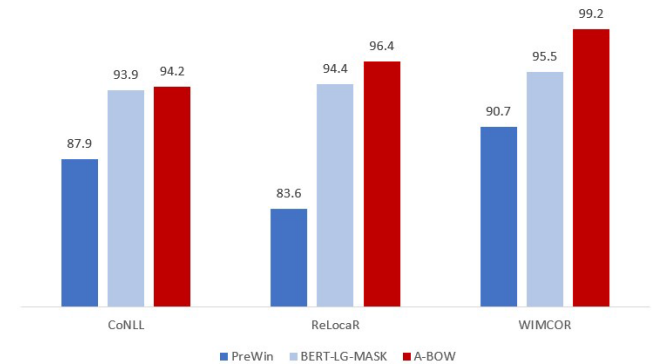


FIGURE 4. The accuracy of PreWin, BERT-LG+MASK and A-BOW on the most recent datasets.

TABLE 3. Results for A-BOW and baseline methods on the different datasets using the same deep learning model. Figures are averaged over 10 runs and the standard deviation.

Method	Dataset	Accuracy	F1 score
A-BOW	SemEval	92.3% ± 0.34	79.8%
Immediate 5	SemEval	82% ± 0.60	29.4%
Immediate 10	SemEval	82.2% ± 0.67	28.8%
Paragraph	SemEval	80.7% ± 0.93	18.5%
A-BOW	ReLocaR	96.4% ± 0.53	96.5%
Immediate 5	ReLocaR	77.3% ± 0.85	76.9%
Immediate 10	ReLocaR	77.1% ± 1.35	77.1%
Paragraph	ReLocaR	73.7% ± 2.70	74.4%
A-BOW	CoNLL	94.2% ± 0.27	91.8%
Immediate 5	CoNLL	83.9% ± 0.01	76.1%
Immediate 10	CoNLL	75.7% ± 0.42	62.0%
Paragraph	CoNLL	34.0% ± 0.01	21.3%
A-BOW	WIMCOR	99.2% ± 0.04	98.4%
Immediate 5	WIMCOR	89.2% ± 0.01	76.5%
Immediate 10	WIMCOR	84.7% ± 0.01	61.8%
Paragraph	WIMCOR	NA	NA

A. A-BOW VS BASELINE METHODS

Table 3 shows that A-BOW method has substantially outperformed the baseline methods with at least +10% increase in accuracy across all datasets, given that both were trained on the same CNNs model. Note that the F1 score is significantly lower in SemEval than the other datasets due to the imbalanced classes of the mentioned dataset. The paragraph method achieved the lowest score in all datasets due to the amount of noise it contains, it prioritizes the length of the sequence over the relevance of the word, as a result, the number of irrelevant words are higher compared to the immediate methods. Unlike the paragraph, our method filters out the words that do not intersect between the A-BOW and the samples of the datasets i.e. segregating the context words that indicate the nature of the location mentioned from the

rest of the words, the size of the sequence is not static as it depends on the number of words that intersect. From the results obtained, the concept of the baseline methods which is: “the closer the word to the PMW the more reliable it can be” does not work as efficiently as A-BOW method, for the reason that, there may be irrelevant words carried along in the sequence which will affect the classification task later on.

B. A-BOW VS PREVIOUS MODELS

In comparison with the earlier Location Metonymy Resolution models, as it is shown in Fig. 3, Our model has outperformed all of the previous models achieving the new SOTA on the SemEval dataset with an accuracy of 92.3%. Note that GYDER and Nastase *et al.* [20], have achieved higher results (85.2% and 86.2 respectively) than PreWin (83.1%) due to their excessive resource usage, whereas, PreWin has substantially reduced the amount of resources consumed to extract features by solely relying on the dataset and a dependency parser (to locate the head modifier of the sentence). Li *et al.* [22], achieved 88.2% using BERT along with a masking approach to exclude the PMW for being passed on as a feature in order to avoid a possible lexical memorization.

Fig. 4 shows our model’s performance in recent datasets, considering that, most of the Metonymy Resolution research was done before the release of these datasets (ReLocaR, CoNLL, WIMCOR), therefore, only two studies have used them so far, which are: Gritta *et al.* [3], and Li *et al.* [22]. Our model has achieved the highest accuracy across all three datasets, the highest accuracy scored is 99.2% on the WIMCOR dataset. Note that Li *et al.* [22] performed an ensemble BERT-LG-MASK with the following accuracies 94.8%, 94.6%, and 95.9% for ReLocaR, CoNLL, and WIMCOR respectively.

TABLE 4. Accuracy of the most recent methods where the source is different from the target. Averaged over 10 runs and the standard deviation.

Source	Target	PreWin	BERT	A-BOW
SemEval	ReLocaR	62.4±2.30	75.2±1.05	85.1±1.01
ReLocaR	SemEval	69.0±3.13	74.8±1.29	80.7±1.85
CoNLL	ReLocaR	82.6±0.87	93.5±0.40	89.3±0.05
CoNLL	SemEval	79.5±0.34	82.5±1.69	90.6±0.06
WIMCOR	ReLocaR	64.1±0.42	64.6±1.05	87.1±3.80
WIMCOR	SemEval	75.6±0.76	78.4±0.97	92.5±1.65

C. CROSS DOMAIN COMAPRISON

As observed in Table 4, the results of the cross-domain experiments has experienced a decline in the overall accuracy primarily due to the different annotation schemes particularly SemEval and ReLocaR datasets and the label distributions; all methods performed poorly in the first two rows compared to the results shown in Fig. 4. In contrast, transferring from CoNLL to either SemEval or ReLocaR shows improvement in all three methods achieving the highest of 93.5% by BERT-

LG-MASK from CoNLL to ReLocaR and 90.6% by A-BOW from CoNLL to SemEval. Despite the fact that WIMCOR contains orders of magnitude more samples, it poorly transfers to either ReLocaR or SemEval with PreWin and BERT-LG-MASK; A-BOW was not hugely affected by the source dataset as the previous two, where it scored its highest accuracy in the sixth row, we hypothesize that the reason of A-BOW being resistant to some degree is the nature of the method itself, where the bigger the bag-of-words is the more insights it carries.

D. ERROR ANALYSIS

To have a further understanding on how our model handles the metonymy resolution task and why it fails when encountering some samples, we conducted an error analysis over a random sample of 300 errors. Table 5 shows some samples of the different types that the classification is prone to falsely categorize. Based on our observations, there are three dominant types of samples that have the potential to mislead the classification task. The first type is “the short samples”: one of the conditions that was set during the noise elimination process (when intersecting the original sample with the A-BOW to find mutual words) is to preserve the samples that have at least two words, in other words, if the intersection between A-BOW and the original sample is only one word, the algorithm switches to the baseline method of Immediate10, the reason of finding wrongly classified samples from this type is that the samples are significantly shorter than being able to give a correct indication of the location’s nature, however, two words can accurately identify the nature of the location in many cases for example: “led by, decided to” for metonymic locations and “located in, situated north” for the literal ones, moreover, increasing the minimum word count of the sample will force the algorithm to switch more frequently to the baseline method, thus, our approach will have less impact on the classification. The second type is “the complex sentence”: the common element of this type is that the sentence contains subsentences within it, hence, the algorithm may choose words from a substance different from the one the PMW is mentioned in as shown in the middle rows of Table 5. The third type is the “immediate10 samples”: as mentioned earlier in the first type, when the word count of the sample is less than two words, Immediate10 replaces our approach which means the algorithm will pick 10 words from each side of the PMW (based on the experiments, Immediate 10 achieved the best results among the other baseline methods we implemented); the downside of the immediate approaches is that they blindly pick the words from the sides of the targeted location regardless of the nature of the word or its relevance to the meaning.

E. DISCUSSION AND CONCLUSION

In this paper, the main novelty is our feature selection approach. We combined the classic bag-of-words method with GloVe word embeddings. By exploiting one of GloVe properties (Similar words or words that frequently co- occur

TABLE 5. Error analysis, the samples are shown after applying the feature selection algorithm where “original label” is the correct label of the sample and “prediction” is what the sample has been classified.

Prediction	Sample	Original Label
Metonymy	But Both	Literal
Metonymy	election most important long ordeal...	Literal
Metonymy	from the company have support ...	Literal
Literal	to since the regional form...	Metonymy
Literal	currently the codex located at...	Metonymy

are represented close to each other), it allowed us to augment the initial BOW with words that may have the same significance of the most repeated ones in the corpus but not frequently mentioned, which resulted in widening the range of the words that can indicate the nature of the PMW. Since our feature selection algorithm depends on obtaining the intersection between the A-BOW and the dataset samples, the more words in the A-BOW, the more words intersecting with the samples, the clearer the class of the targeted location. Our approach was tested on four datasets (SemEval, ReLocar, CoNLL, WIMCOR), we trained our model on a minimalist CNNs. The results show a remarkable improvement in the performance of *Location, Metonymy Resolution classification task*, we achieved **SOTA** across all datasets that our approach was tested on.

The model proved that it is not necessary to use a significant amount of external resources to boost the performance of the classification task (we only used GloVe and the dataset). It also showed that we can disregard the majority of the sentence and only preserve the words that can reflect the category of the PMW. Our method was inspired by PreWin method [3], as it changed the conventional way of looking at the task, they considerably reduced the external resources yet still achieved competitive results. However, their method has some drawbacks, the main idea of PreWin method is to find and extract the head modifier of the sentence and the four words that come after it, the head modifier is a good starting point to narrow the search for context words but indiscriminately taking the four following words can still result on carrying noise to the learning phase.

In the future, we plan to combine PreWin and A-BOW by finding the intersection using A-BOW inside the predicate window only and not the entire sentence or paragraph, this can help us overcome some compound sentences that may give us a faulty observation on the category of the location e.g. “Brazil that is famous for winning the football world cup 5 times is the largest country in South America”, clearly, the words “famous, winning, football, cup” can be a strong indication that the entity “Brazil” is used metonymically, yet, the main sentence is talking about an actual location. By implementing the head modifier as in PreWin, we can skip the subsentence “that is famous for winning the football world cup 5 times” and narrow it down to “Brazil is the largest country in South America”. Other future work may involve testing NER taggers or geographical parsers that inte-

grate our model to evaluate the location tags after being able to distinguish between literal and metonymic use of locations.

ACKNOWLEDGMENT

The authors would like to express their great appreciation to Prof. Trevor Cohn for his valuable and constructive suggestions during the planning of this research work.

REFERENCES

- [1] G. Lakoff and M. Johnson, “The metaphorical structure of the human conceptual system,” *Cognit. Sci.*, vol. 4, no. 2, pp. 195–208, Apr. 1980.
- [2] B. R. Monteiro, C. A. Davis, and F. Fonseca, “A survey on the geographic scope of textual documents,” *Comput. Geosci.*, vol. 96, pp. 23–34, Nov. 2016.
- [3] M. Gritta, M. T. Pilehvar, N. Limsopatham, and N. Collier, “Vancouver welcomes you! Minimalist location metonymy resolution,” in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2017, pp. 1248–1259.
- [4] F. M. Ferraro, *Toward Improving Automated Classification Metonymy Text Corpora*. Rochester, NY, USA: Honors Bachelor of Science Thesis, 2011.
- [5] E. Shutova, “Sense-based interpretation of logical metonymy using a statistical method,” in *Proc. ACL-IJCNLP Student Res. Workshop (ACL-IJCNLP)*, 2009, pp. 1–9.
- [6] E. Shutova and S. Teufel, “Logical metonymy: Discovering classes of meanings,” in *Proc. CogSci Workshop Semantic Space Models*, 2009, pp. 29–34.
- [7] M. Lapata and A. Lascarides, “A probabilistic account of logical metonymy,” *Comput. Linguistics*, vol. 29, no. 2, pp. 261–315, Jun. 2003.
- [8] D. Fass, “Met*: A method for discriminating metonymy and metaphor by computer,” *Comput. Linguistics*, vol. 17, no. 1, pp. 49–90, 1991.
- [9] J. Pustejovsky, “The generative lexicon,” *Comput. Linguistics*, vol. 17, no. 4, pp. 409–441, 1991.
- [10] J. R. Hobbs, M. E. Stickel, D. E. Appelt, and P. Martin, “Interpretation as abduction,” *Artif. Intell.*, vol. 63, nos. 1–2, pp. 69–142, Oct. 1993.
- [11] K. Markert and U. Hahn, “Understanding metonymies in discourse,” *Artif. Intell.*, vol. 135, nos. 1–2, pp. 145–198, 2002.
- [12] K. Markert and M. Nissim. (2005). *Annotation Scheme for Metonymies ASI*. Rapport Technique, Université de Leeds, Université d’Edinburgh. Disponible Sur. [Online]. Available: <http://www.comp.leeds.ac.uk/markert/Papers/index.html>
- [13] M. Nissim and K. Markert, “Syntactic features and word similarity for supervised metonymy resolution,” in *Proc. 41st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2003, pp. 56–63.
- [14] K. Markert and M. Nissim, “Towards a corpus annotated for metonymies: The case of location names,” in *Proc. LREC*, May 2002, pp. 1–8.
- [15] K. Markert and M. Nissim, “SemEval-2007 task 08: Metonymy resolution at SemEval-2007,” in *Proc. 4th Int. Workshop Semantic Evaluations (SemEval)*, 2007, pp. 36–41.
- [16] K. Markert and M. Nissim, “Data and models for metonymy resolution,” *Lang. Resour. Eval.*, vol. 43, no. 2, pp. 123–138, Jun. 2009.
- [17] R. Farkas, E. Simon, G. Szarvas, and D. Varga, “GYDER: Maxent metonymy resolution,” in *Proc. 4th Int. Workshop Semantic Evaluations (SemEval)*, 2007, pp. 161–164.
- [18] C. Brun, M. Ehrmann, and G. Jacquet, “XRCE-M: A hybrid system for named entity metonymy resolution,” in *Proc. 4th Int. Workshop Semantic Eval. (SemEval)*, 2007, pp. 488–491.
- [19] V. Nastase and M. Strube, “Combining collocations, lexical and encyclopedic knowledge for metonymy resolution,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, vol. 2, 2009, pp. 910–918.
- [20] V. Nastase, A. Judea, K. Markert, and M. Strube, “Local and global context for supervised and unsupervised metonymy resolution,” in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2012, pp. 183–193.
- [21] K. A. Mathews and M. Strube, “A large harvested corpus of location metonymy,” in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 5678–5687.
- [22] H. Li, M. Vasardani, M. Tomko, and T. Baldwin, “Target word masking for location metonymy resolution,” 2020, *arXiv:2010.16097*.
- [23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

- [24] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014, *arXiv:1404.2188*.
- [25] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.
- [26] J. Liu, W. C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2017, pp. 115–124.
- [27] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," 2014, *arXiv:1412.1058*.
- [28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [30] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," 2016, *arXiv:1606.01781*.
- [31] A. B. Duque, L. L. J. Santos, D. Macêdo, and C. Zanchettin, "Squeezed very deep convolutional neural networks for text classification," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2019, pp. 193–207.
- [32] H. T. Le, C. Cerisara, and A. Denis, "Do convolutional networks need to be deep for text classification," in *Proc. Workshops 32nd AAAI Conf. Artif. Intell.*, 2018.
- [33] X. Zhang, J. Zhao, and Y. Lecun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, Sep. 2015, pp. 649–657.
- [34] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.
- [35] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA, USA: Sage, 2018.
- [36] J. Carletta, "Assessing agreement on classification tasks: The Kappa statistic," 1996, *arXiv:cmp-lg/9602004*.
- [37] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *Proc. 6th Int. Semantic Web Conf. (ISWC), 2nd Asian Semantic Web Conf. (ASWC)*, 2007, pp. 715–728.
- [38] J. Lehmann, "DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.
- [39] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
- [40] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Comput. Surveys*, vol. 54, no. 3, pp. 1–40, Apr. 2022.
- [41] S. Du and H. Wang, "Addressing syntax-based semantic complementation: Incorporating entity and soft dependency constraints into metonymy resolution," *Future Internet*, vol. 14, no. 3, p. 85, Mar. 2022.



ROHANA BINTI MAHMUD received the Ph.D. degree from The University of Manchester, U.K. She is currently a Senior Lecturer with the Department of Artificial Intelligence (AI), Faculty Computer Science and Information Technology, Universiti Malaya (UM), Malaysia. She has been teaching various courses in AI and computer sciences disciplines and her expertise's are in natural language processing (NLP), expert system, and machine learning.



SAMEEM BINTI ABDUL KAREEM received the B.Sc. degree (Hons.) in mathematics from the University of Malaya, Kuala Lumpur, in 1986, the M.Sc. degree in computing from the University of Wales, Cardiff (now known as the University of Cardiff), in 1992, and the Ph.D. degree from the University of Malaya, in 2002. She started her career as a Lecturer with the Institute of Preparatory Studies, ITM (now known as UiTM), in 1986. She was the Dean of the Faculty of Computer Science, from 2019 to 2021. She is currently an Academic with the Department of Artificial Intelligence, University of Malaya. She has published more than 100 WOS journal articles and conference proceedings. Her research interests include artificial intelligence in medicine, machine learning, data analytics/mining, image processing, and biomedical informatics.



ASSAAD OUSSAMA ZEGHINA received the master's degree in machine learning and data science from the University of Paris, France, in 2021. He is currently pursuing the Ph.D. degree in graph mining with the University of Strasbourg. His research interests include deep learning, graph neural networks, and computer vision.



YOUNES SAADI received the Ph.D. degree in computer science from University Malaya, Malaysia. He is currently working as a Senior Data Scientist in food technology, where he applies several techniques to optimize farming yielding, quality, storage, and sales. He published many research papers in journals and conferences. His research interests include optimization algorithms, recommendation algorithms, and time series forecasting.

...



MUHAMMAD ELYAS MEGUELLATI received the bachelor's degree in information system from the University of Batna 2, Algeria, in 2018. He is currently pursuing the M.Sc. degree in applied computing with the University of Malaya. His interests include natural language processing, computer vision, and machine learning.