Zero-shot Text Classification via Knowledge Graph Embedding for Social Media Data

Qi Chen, Wei Wang, Kaizhu Huang, and Frans Coenen

Abstract—**The idea of 'citizen sensing' and 'human as sensors' is crucial for social Internet of Things, an integral part of cyber-physical-social systems (CPSS). Social media data, which can be easily collected from the social world, has become a valuable resource for research in many different disciplines, e.g. crisis/disaster assessment, social event detection, or the recent COVID-19 analysis. Useful information, or knowledge derived from social data, could better serve the public if it could be processed and analyzed in more efficient and reliable ways. Advances in deep neural networks have significantly improved the performance of many social media analysis tasks. However, deep learning models typically require a large amount of labeled data for model training, while most CPSS data is not labeled, making it impractical to build effective learning models using traditional approaches. In addition, the current state-of-the-art, pre-trained Natural Language Processing (NLP) models do not make use of existing knowledge graphs, thus often leading to unsatisfactory performance in real-world applications. To address the issues, we propose a new zero-shot learning method which makes effective use of existing knowledge graphs for the classification of very large amounts of social text data. Experiments were performed on a large, real-world tweet dataset related to COVID-19, the evaluation results show that the proposed method significantly outperforms six baseline models implemented with state-of-the-art deep learning models for NLP.**

Index Terms—**Natural language processing, knowledge graph, zero-shot learning, internet of things, social media data analysis.**

I. Introduction

With the rise of smart devices and technologies, the Internet of Things (IoT), mobile social networks and cloud computing, 'human as sensors' or 'citizen sensing' [1] has become a popular phenomenon for which humans are not only the data users, but also the data providers. It allows the general public to collect, analyze, report and disseminate information, enabling them to better perceive and understand the world. Meanwhile, it is crucial for the development of social IoT [2], an integral part of the Cyber-Physical-Social systems (CPSS) [3], [4]. Enormous amount of social media data can be collected and further processed and analyzed in various downstream tasks which may have great influence on human society. For example, users may post real-time traffic information on Twitter, which facilitates traffic event detection [5]. Other examples include reports of injured or missing people, infrastructure damage, and warnings and cautions; which all help crisis/disaster assessment and emergency response [6], [7].

To extract useful information and knowledge from social media data, Natural Language Processing (NLP) techniques are usually adopted. Recently, Deep Neural Networks (DNNs) have shown impressive performance in NLP, image processing, and many other data mining tasks. Under the traditional supervised learning paradigm, DNNs have become unbeatable in terms of classification performance, provided that there are sufficiently large amounts of well labeled examples. Example application domains include vehicle identification from images, document classification, and neural machine translation. However, they usually break down when there is not sufficient labeled data. The ability to transfer the knowledge gained while solving one problem and applying it to a different but related problem (referred to as transfer learning) can alleviate this issue. One notable example of using transfer learning in NLP, so far, is to pre-train representations on a large unlabelled text corpus and then adapt the trained representation to a supervised target task. A number of pre-trained models have been developed very recently, e.g. word2vec [8], GloVe [9] and Bidirectional Encoder Representations from Transformers (BERT) [10], that have been applied to various tasks, e.g. image caption generation [11], sentiment analysis from social media [12], and text classification in smart city application [6]. Besides the use of pre-trained models, the research community has also shown great interest in other forms of transfer learning, e.g. domain adaptation [13], multi-task learning [14], zero-shot learning [15], etc. Particularly, zero-shot learning requires a classifier to recognize samples from classes that were not observed during training. Such characteristic makes it especially suitable for processing and analyzing social media data, as social media data is mostly unlabeled and it is difficult to label a good amount of data representative of various classes.

With the unprecedented volume of data generated from CPSS, the use of advanced graph-based methods, e.g. graph embedding and graph neural network, to model the relationship between data items has become a promising research direction. Although many studies have been conducted, research on how to efficiently and effectively exploit the power of graph-based methods, given the overwhelming amount of CPSS data, is still at an early stage. Recently, research on how to effectively utilize existing, quality knowledge bases within DNNs has attracted significant attention [16]. The knowledge stored in many existing knowledge bases and knowledge graphs represents facts and human wisdom accumulated over centuries. Including such knowledge in learning systems has

Qi Chen, Wei Wang and Kaizhu Huang are with the School of Advanced Technology, Xi'an Jiaotong Liverpool University, China. E-mail: {qi.chen,wei.wang03,kaizhu.huang}@xjtlu.edu.cn.

Frans Coenen is with Department of Computer Science, University of Liverpool, Liverpool, UK. e-mail: Coenen@liverpool.ac.uk

Manuscript received xxx, 2021; revised .

great potential. On one hand, systems do not need to learn existing knowledge from scratch; on the other, mistakes made in classification previously can be avoided to a great extent. Embedding has emerged as an important approach to prediction, inference, data mining, and information retrieval. Graph embedding algorithms that represent the hierarchical structure of a knowledge base in the format of vectors have been increasingly investigated. By transferring the rich structural information from knowledge bases to learning systems, better prediction, classification and recommendation performance could be anticipated. However, the convergence of deep learning and knowledge graph embedding is a challenging research topic that has not been extensively studied.

For the issues discussed above, we propose a new zero-shot learning method which exploits the use of existing knowledge graphs for the classification of large amounts of social text data (i.e. Twitter messages related to COVID-19) without training data. Following the key ideas in zero-short learning, the proposed method does not explicitly define the class labels. The pre-trained sentence based BERT model (S-BERT) [17] is first used to represent tweet messages in the embedding space to be further matched with classes. As the purpose of S-BERT is to learn a sentence-level representation, while most class labels contain only one or few words, S-BERT embedding may not be as semantically consistent as word-level embedding methods. To address this problem, we construct a knowledge graph embedding model for label representation with a comprehensive knowledge graph named ConceptNet [18]. The sentence embedding is then projected to the knowledge graph through the least-squares linear projection. The proposed model is referred to as the S-BERT-KG model. We apply the model to COVID-19 related tweet classification without any labeled data for training. To our best knowledge, this is the first work that takes the zero-shot learning architecture for tweet classification. Experimental results demonstrated that the proposed S-BERT-KG model is able to gain significant improvement over other baseline models and produce reasonable prediction accuracy without any labeled data.

The rest of the paper is organized as follows. We review some representative work on social media data classification, zero-shot text classification, graph neural networks, and graph embedding in Section II. In Section III, we describe the S-BERT-KG architecture, knowledge graph embedding, and the zero-shot text classification procedure in detail. Section IV describes the experiments conducted on a large Twitter dataset related to COVID-19 and presents the evaluation results compared with baseline models. Finally, we conclude the paper and discuss some future research directions in Section V.

II. Related Work

In this section, we review some representative work on social media data classification, zero-shot text classification, and discuss some latest applications using graph neural networks and graph embedding.

Social media data classification: As of May 2020, there are around 500 million tweets posted on Twitter each day. Social media "senses" nearly everything happening around the world, and the produced data has become a valuable source for research in different disciplines. In contrast to data collected from the physical world, social media data has some attractive features. For example, it covers far more areas and topics, can be collected at low cost, and enjoys high-level semantics understandable to human users. With the ideas of 'citizen sensing' and 'human as sensors', a number of real-world applications have been developed, e.g. traffic event detection [5], spammer detection [19], and natural disaster assessment [6], [7]. Various NLP techniques have been applied to extract useful information from short tweet messages. For instance, the work in [7] proposed a large word2vec embedding that was trained on 52 million crisis-related tweets and used to classifies crisis tweets using Support Vector Machines, Naive Bayes and Random Forest methods. The study in [6] designed a framework based on the Convolutional Neural Network (CNN), which used convolution and pooling operations to capture important information to identify useful or crisis-related tweets. The work in [19] developed a collaborative neural network spammer detection mechanism, which fused multi-source information by collaboratively encoding long-term behavioral and semantic patterns. The work in [5] implemented CNN and Recurrent Neural Network (RNN) deployed on the top of word-embedding models for detecting traffic events. Recently, Twitter has seen a massive surge in the daily traffic related to the COVID-19 pandemic outbreak. A number of studies has collected a large amount COVID-19 related social media data [20] and applied topic modeling techniques [21] to analyze topic trends. In short of labeled data, studies exploiting supervised or semi-supervised NLP models to automatically classify tweets are still very limited. In this paper, we apply state-of-the-art NLP techniques to extract knowledge from COVID-19 related messages on social media in an unsupervised manner.

Zero-shot text classification: To extract knowledge from Cyber-Physical-Social (CPS) data, traditional supervised learning paradigm needs sufficient labelled data for model training, while most CPS data is not labeled. Recently, the research focus has shifted towards learning from a large amount of unlabeled data, and has achieved remarkable progress in domain adaptation [13], few-shot learning [22], zero-shot learning [15], or more generally, transfer learning [23]. Bidirectional Encoder Representations from Transformers (BERT) [10] and its variants (e.g, S-BERT [17], RoBERTa [24], BART [25]) are pre-trained language models that achieved state-of-the-art results in various text classification tasks, e.g., machine translation, question answering and language inference. While in the zero-shot classification scenario, a classifier is required to work on labels that it is not explicitly trained with. In the computer vision domain, one common approach for zero-shot learning is to use existing feature extractors to represent images and any possible label names in their corresponding embedding space [26]. Some annotated data might be used to align the image and label embedding. The framework allows any label (seen or unseen during training) and any image to be embedded in the same latent space to measure their similarity. In the text domain, one single NLP model can be utilized to embed both data and any class names into the same embedding space without an alignment step. In particular, [27] concatenated both the sentence data and class names as model input, and treated zero-shot learning as a binary classification task; [28] took both sentence and label names into the BERT model and considered zero-shot text classification as a Natural Language Inference (NLI) task. Inspired by the work in [17] and [28], we propose a zero-shot text classification architecture for social media data analysis. The proposed method directly makes use of models pre-trained with NLI tasks, thus does not need any labeled data for model training.

Graph neural network and graph embedding: The growing number of connected IoT equipments can be described with graphs, e.g. vertices denote sensors while edges denote the connection between them. Understanding the structure of such complex and ubiquitous IoT networks remains a challenging task. Graph Neural Networks (GNNs) [29], a class of deep learning method designed to perform inference on data described by graphs, motivated researchers to model IoT data based on the internal relationships between different sensors. One typical application of GNNs in IoT is traffic prediction, where the traffic network can be modeled as a spatial-temporal graph; the nodes are loop sensors installed on roads, and the edges represent the intersections or road segments connecting these sensors. The work in [30] and [31] utilized Graph Convolutional Networks (GCNs) to capture spatial dependency, leveraged recurrent neural networks (RNNs) for modeling temporal dynamics, and attained state-of-the-art performance in the traffic prediction task. With the enormous amount of cyber data available on the Internet, significant efforts from industry and academia have been made into constructing knowledge bases [16], e.g. DBpedia [32] and ConceptNet [18]. To represent the hierarchical structure of a knowledge base in the format of vectors, graph embedding methods have been increasingly investigated, e.g. DeepWalk [33], Node2vec [34], SDNE [35], etc. Meanwhile, recent studies [36], [37] leveraged these external knowledge graphs to further enhance language representation by integrating the rich structured knowledge facts into NLP model input. In this paper, we further explore the knowledge graph embedding technique in the zero-shot text classification scenario for better natural language understanding.

III. Zero-shot Learning with Knowledge Graph

In this section we describe the S-BERT-KG model in terms of a tweet classification task for extracting informative tweets. Let X be the set of tweets to be categorized and L be the set of possible class names (or labels) for X. The goal is to represent both tweets X and label names L in the same embedding space, so as to classify any tweets by measuring the similarity between tweet embeddings (usually in a form of a sentence) and label embeddings, without using any labeled data. It should be noted that the set of possible label names L are not explicitly defined in the zero-shot scenario, which can be any object in a given knowledge graph or words in a given vocabulary. As it is impractical to evaluate the model on all possible labels in a vocabulary, we need to specify a possible label set, i.e., seven representative classes of COVID-19 tweets in this study.

The overall architecture of S-BERT-KG is shown in Figure 1. A pre-trained Sentence BERT (S-BERT) [17] is firstly used to embed sentence $u = f(x, \theta_f)$ and label $v = f(y, \theta_f)$ where $x \in X$ and $y \in L$. An external knowledge graph (i.e. ConceptNet) is used to construct a knowledge graph embedding space with the retrofitting [38] method. Sentence embedding and label embedding are then projected to the knowledge graph embedding space via a learned projection matrix P to determine if a tweet belongs to a specific class(es) based on their cosine similarity. In the following, we firstly introduce the relevant BERT models, then present the method used for knowledge graph embedding, and finally show the zero-shot text classification process.

A. Bidirectional Encoder Representations from Transformers (BERT) and Sentence BERT (S-BERT)

As an important component of modern NLP tasks, pre-trained word embeddings (such as word2vec [39] and GloVe [9]) can substantially improve the performance of NLP tasks compared to the embeddings learned from scratch. For each word in the vocabulary, these context-free models generate a single word embedding representation, despite the fact that the meaning of such words may vary in different scenarios. While the contextual models, such as OpenAI GPT [40], ELMo [41], and BERT [10], can generate a representation for each word based on the contexts (i.e. surrounding words in a sentence). These contextual models usually contain more hidden layers and require a lot of unlabeled data for training. When these pre-trained contextual models are applied in domain-specific tasks, fine-tuning with a small amount of labeled data is usually sufficient.

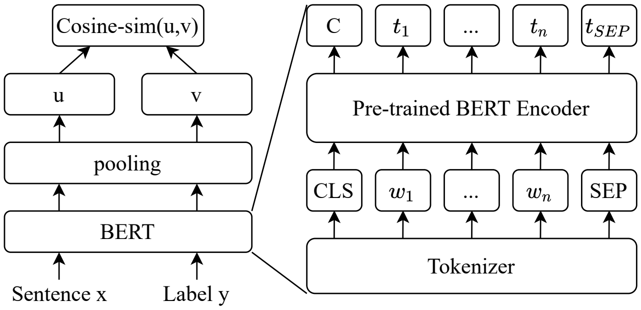


Fig. 2. Sentence-BERT (S-BERT) architecture

BERT has achieved state-of-the-art performance in a number of NLP tasks, e.g. text classification, question answering, and language inference. However, for sentence-pair regression tasks such as language inference, BERT requires that both sentences are fed into the network simultaneously, which demands significant computation at resources especially when the number of sentences is large. The recent Sentence-BERT architecture [17], that uses the Siamese structure to derive semantically meaningful sentence embeddings, reduces the effort for finding the most similar pair while maintaining accuracy. For zero-shot classification purposes, instead of taking sentence pairs, the S-BERT model takes a Sentence x and a Label y into the model to measure their similarity.

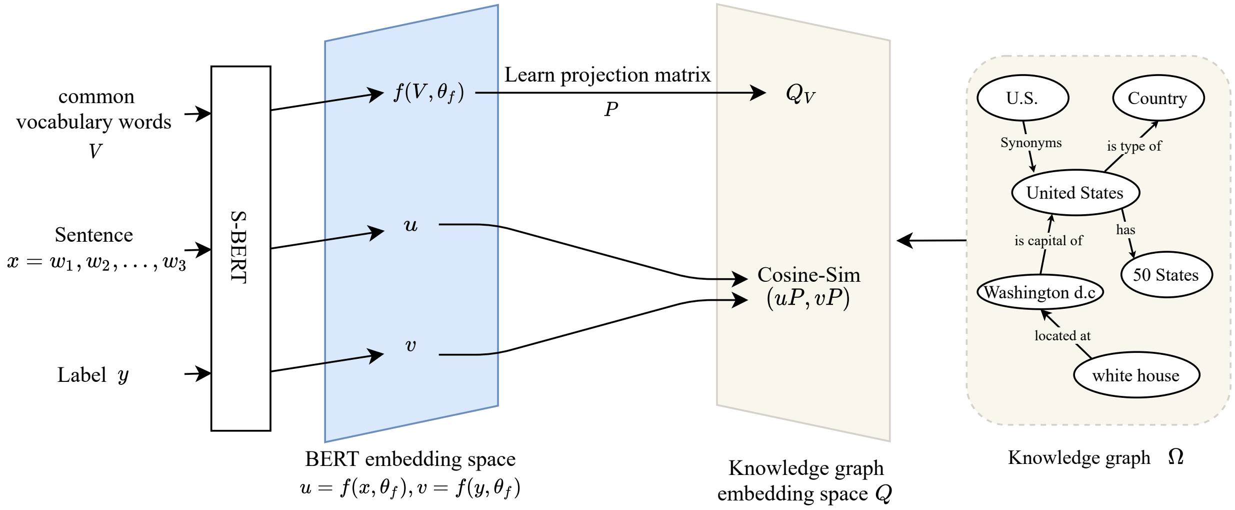


Fig. 1. S-BERT with Knowledge Graph embedding (S-BERT-KG) architecture

The architecture of the S-BERT model is illustrated in Figure 2.

As shown in Figure 2, a tokenizer is used to split the tweet input into tokens $\{w_1, w_2, ..., w_n\}$, add special tokens (e.g., [CLS] and [SEP]) and convert these tokens into indices of the tokenizer vocabulary. The output hidden state C is used as the aggregated sequence representation for the classification tasks and $\{t_1, t_2, ..., t_n\}$ represents the corresponding word embedding vectors for tokens $\{w_1, w_2, ..., w_n\}$. S-BERT contains an additional pooling layer to evaluate models under different pooling strategies by using: 1) the output of CLS-token (C); 2) the mean of all hidden states, or 3) the maximum of all output vectors. It is trained with SNLI [42], Multi-NLI [43], and STS [44] datasets so as to provide similar sentence embeddings for semantically similar sentences [17]. On the basis of this idea, we could directly take the pre-trained model and classify a sentence x to a label $\hat{y}$ with Eq. 1, given all possible label names L.

$$\begin{aligned}
\hat{y} &= \arg\max_{y \in L} \cos(u, v) \\
u &= f(x, \theta_f), x \in X \\
v &= f(y, \theta_f), y \in L
\end{aligned} \tag{1}$$

where *cos* is the cosine similarity, f is the function represented by S-BERT, θ_f denotes the pre-trained parameters of f, X and L represent the set of tweets and possible class labels respectively.

B. Knowledge Graph Embedding

One issue using S-BERT for zero-shot text classification is that S-BERT is trained to learn effective sentence-level representations, but may not generate semantically consistent single word label representations as other word embedding methods do (e.g. word2vec and GloVe). Also, all pre-trained language models (e.g. BERT, word2vec, and GloVe) lack common-sense or domain-specific knowledge, which usually results in unsatisfactory performance for short message representation. To address these two issues, we construct a knowledge graph embedding model based on ConceptNet [18] for zero-shot tweet classification.

ConceptNet [18] is a knowledge graph that connects words and phrases of natural language with labelled edges. Its knowledge is collected from multiple sources that include expert generated resources and crowd-sourcing. It aims to represent the common sense knowledge involved in understanding language and to improve natural language applications by enabling applications to better understand the meaning behind the words. ConceptNet represents relations and words (e.g. shown in Fig. 1) as triples, e.g. (Washington d.c., capitalOf, United States), (United States, has, 50 States) and (United States, typeOf, Country).

By adding the rich graph structure information from ConceptNet into common NLP techniques, particularly, word embedding methods (i.e. word2vec and GloVe), we could construct a semantic space that is potentially more effective than distributional semantics in terms of better prediction, classification and recommendation performance. Retrofitting [38] is a process that adjusts an existing matrix of word embeddings with a knowledge graph. It calculates new word vectors q_i with the objective of staying close to both their original values in other word embedding vectors (i.e. word2vec and GloVe) $\hat{q}_i$, and their neighbors q_j in the knowledge graph with edges $(i, j) \in E$. The knowledge graph embedding Q can be computed by minimizing the following objective function:

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \left\| q_i - \hat{q}_i \right\|^2 + \sum_{(i,j) \in E} \beta_{ij} \left\| q_i - q_j \right\|^2 \right] \tag{2}$$

where $Q = (q_1, q_2, ..., q_n)$ is the learned knowledge graph embedding matrix, α and β values control the weights of word embedding and knowledge graph. The word vectors in Q are firstly initialized to be equal to the vectors in $\hat{Q}$ and then retrofitted. An iterative updating method [45] is used to calculate Q that converges in just a few iterations.

C. Embedding Alignment and Classification Process

We have embedded tweet and label representations into the S-BERT embedding space, and constructed a knowledge graph embedding for all possible labels. Next, we need to learn a linear projection function that can align the tweet and label embeddings. This allows one to embed any tweets and labels into the same knowledge embedding space to measure their similarity. For text data, this process can be done by aligning the representations of the same vocabulary words in different embedding space. We take the top K frequently used English words from the vocabulary of the S-BERT model, and learn a projection matrix P with least-squares linear projection. The process maps the word embeddings from the S-BERT embedding space to the knowledge graph embedding space. For label representations, whether using the representation after projection $f(y, \theta_f)P$, or directly deploying the embedding from knowledge graph embedding space Q_y, leads to similar prediction performance. After this embedding alignment process, the equation to classify a given tweet becomes:

$$\hat{y} = \arg\max_{y \in L} \cos(f(x, \theta_f)P, f(y, \theta_f)P) \tag{3}$$

The embedding alignment and the zero-shot text classification of using the S-BERT-KG method are then detailed in Algorithms 1 and 2 respectively. The purpose of the embedding alignment process is to learn a projection matrix P that maps the S-BERT representation to the knowledge graph embedding space. In Algorithm 1, we first learn a knowledge graph embedding using ConceptNet (Line 1), then obtain the representation of the most common vocabulary words from S-BERT and the knowledge graph embedding (Lines 2-5), and finally learn a projection matrix (Line 6). For zero-shot text classification (as shown in Algorithm 2), we use the tweet representations (Lines 6-7) and the label representations (Lines 2-3) to generate label prediction (Line 8) without any training process.

Algorithm 1 S-BERT-KG embedding alignment

```
Input: pre-trained S-BERT model parameters $\theta_f$, GloVe word embedding $\hat{Q}$, knowledge graph $\Omega$
Output: Projection Matrix $P$
1: Calculate knowledge graph embedding $Q$ with GloVe word embedding $\hat{Q}$ and knowledge graph $\Omega$ using Eq. 2;
2: Initialize S-BERT model with pre-trained parameters $\theta_f$;
3: Take $K$ most common vocabulary words $V$ from the S-BERT model;
4: Obtain knowledge graph embeddings for words $V$: $Q_V$;
5: Obtain their S-BERT embeddings for words $V$: $f(V, \theta_f)$;
6: Learn a least-squares linear projection matrix $P$ from $f(V, \theta_f)$ to $Q_V$.
```

Algorithm 2 S-BERT-KG zero-shot text classification

```
Input: pre-trained S-BERT model parameters $\theta_f$, unlabeled tweet dataset $X$, possible label names $L$, projection matrix $P$
Output: label predictions $\hat{Y}$.
1: for each label $y$ in $L$ do
2:     Represent each label $y$ with S-BERT embedding: $f(y, \theta_f)$;
3:     Project $f(y, \theta_f)$ into the knowledge graph embedding space with projection matrix $P$, which gives $f(y, \theta_f)P$;
4: end for
5: for each tweet $x$ in $X$ do
6:     Represent each tweet $x$ from $X$ with S-BERT embedding: $f(x, \theta_f)$;
7:     Project $f(x, \theta_f)$ into the knowledge graph embedding space with projection matrix $P$, which gives $f(x, \theta_f)P$;
8:     Generate label prediction $\hat{y}$ for tweet $x$ with Eq. 3;
9: end for
```

IV. Experiments

In this section, we first describe the large unlabeled COVID-19 Twitter dataset we collected from Twitter and the two small labeled datasets we manually created for the evaluation purpose. We then report the evaluation results and compare the performance of S-BERT-KG with several state-of-the-art deep learning methods.

A. Dataset

TABLE I
Distributions of the COVID twitter datasets

Classes	Multi-class Dataset (D1)	Multi-label Dataset (D2)
Advice	137	155
China	449	554
Mask	225	272
News	309	408
Transportation	46	57
USA	476	596
Vaccine	96	115
total	1,738	1,941

A number of recent studies have been conducted to process COVID-19 related tweets, e.g. sentiment analysis and topic modeling. However, research that focuses on extracting informative tweets and classifying them to meaningful classes has rarely been found due to the lack of a labeled dataset. Moreover, we have not spotted any public COVID-19 related twitter datasets that are labeled with meaningful categories. We collected all COVID-19 related tweets with tweet IDs provided by GeoCoV19 [20], which contain more than 524 million multilingual tweets from Feb 1st to May 1st 2020. We took one week to collect all the tweets with IDs in GeoCoV19 using the Twitter API, and formed a 12.8GB dataset. So far, we have not seen any notable research on zero-shot learning with unlabeled Twitter datasets. In this study, we only used the tweets written in English, and preprocessed the tweet text to lower case and removed punctuation, question marks and/or URLs.

To evaluate the performance of the proposed model using standard evaluation metrics for supervised learning, we manually labeled some tweets and constructed two small datasets. Hashtags represent keywords or topics in a tweet message. To define meaningful categories for the datasets, we collected frequently used hashtags from the COVID-19 twitter dataset (e.g., #COVID19, #socialdistancing, #wuhan, #covid19_US, #vaccine, #n95, etc.). As it was impractical to define very detailed categories (e.g., COVID19 in US, COVID19 in UK, COVID19 in China, COVID19 in Italy, COVID19 in Spain, etc.) to label all these tweets, we manually examined the hashtags and selected seven trending topics related to the COVID-19 pandemic. It should be noted that the selected seven labels are exploited only for the evaluation purposes. We could select any word from a knowledge graph vocabulary as a label for zero-shot text classification. For each day between Feb 1st and May 1st, we sampled a number of tweets and categorized them into the following seven classes.

1) **Advice**: Stay at home, wash hands, wear mask or social distancing.
2) **China**: Wuhan, China Coronavirus Updates, China news, or other tweets related to China.
3) **Mask**: Mask shortage, wear mask, mask types, etc.
4) **News**: Coronavirus updates, news, rules, etc.
5) **Transportation**: Flights, traffic, traveling, etc.
6) **USA**: U.S. Coronavirus Updates, U.S. news, or other tweets related to the United States.
7) **Vaccine**: Vaccine news, vaccine progress, vaccine injection, etc.

One labeled dataset is for evaluation of multi-class classification, and the other for multi-label classification as one tweet may be related to multiple topics. After removing non-informative and duplicate tweets, the two datasets (D1 and D2) contained 1,738 and 1,941 labeled tweets respectively. D2 contains all the tweets in D1 and an additional 203 tweets with multiple labels. The detailed distribution of the datasets is shown in Table I.

For the experiments we used the ConceptNet 5.7, which contains over 21 million edges and over 8 million nodes (around 1.5 million English nodes). We constructed a sub knowledge graph from ConceptNet using its API with all the vocabulary words that appearing in the COVID-19 Twitter dataset. However, we found that the constructed sub-graph embedding did not provide any improvement over the original version of ConceptNet Numberbatch[1]. Therefore, we utilized the original version in our proposed architecture for better reproducibility.

B. Baseline Models

We re-implemented six baseline models for zero-shot multi-class and multi-label classification, and compared their performance with the proposed S-BERT-KG model.

- **GloVe-AVG**: We used GloVe [9] word embedding to represent each word in a sentence and all the possible label names. The averaged embedding vector was used to represent the sentence and further measure distance with each label.
- **BERT-CLS and BERT-AVG**: We used the last hidden state output of the standard BERT [10] model to represent sentence embedding and label embedding. The output of the CLS token was used for the BERT-CLS version; and the averaged last hidden state output was used for the BERT-AVG version.
- **S-BERT**: We used Sentence-BERT [17] pre-trained with SNLI, MultiSNL and STS datasets to represent sentence and label embeddings for zero-shot classification.
- **S-BERT-GloVe**: Besides using the S-BERT model, We also learned a function to project S-BERT embedding into the GloVE embedding space.
- **BART-NLI**: The study presented in [28] showed that zero-shot text classification can be modeled in a natural language inference architecture, where the hypothesis is constructed by associating a label, e.g., “news”, with the pre-defined problem “The text is about ?”. Given a sentence, the model is trained to determine whether the hypothesis is true. A recent BART model [25], pre-trained with the SNLI dataset, was used in our zero-shot tweet classification task for comparison.

In this paper, we chose a traditional word embedding model (GloVe-AVG) and five deep learning based models for performance evaluation. We show that the zero-shot text classification task can be modeled as a natural language inference problem by comparing BERT-CLS and BERT-AVG with S-BERT. We compare S-BERT and S-BERT-GloVe with the proposed S-BERT-KG to demonstrate the effectiveness of incorporating an external knowledge graph for better language understanding. We also show the performance of the proposed model against the state-of-the-art BART-NLI model.

C. Setup

The GloVe [9] word embedding vectors used in the experiments contain 400K uncased words in the vocabulary and were pre-trained with 6 billion tokens[2]. We use the GloVe word vector with 300 dimensions for all experiments.

For the BERT-CLS and BERT-AVG models, we used the pre-trained BERT-uncased-base model, which contained 12 transformer blocks and 768 dimensions in the hidden units. For the S-BERT, S-BERT-GloVe and the proposed S-BERT-KG models, we took the large S-BERT model pre-trained with the SNLI and STS datasets as described in [17]. The “mean” strategy for the pooling layer was applied.

For the S-BERT-GloVe and S-BERT-KG, we selected 20,000 most frequently used words from the BERT vocabulary to learn the projection matrix P. For BART-NLI, we engaged the large BART model pre-trained with the SNLI dataset. We applied the transformers[3] and sentence-transformers[4] libraries to implement all the models; all pre-trained models used in our study can be downloaded via these two libraries.

[1] https://github.com/commonsense/conceptnet-numberbatch

[2] https://nlp.stanford.edu/projects/glove/

[3] https://huggingface.co/transformers/

[4] https://www.sbert.net/

The experiments were run using PyTorch 1.7.1, Tensorflow 2.4, Python 3.6, and Windows 10 running on a desktop computer with an i7-9700F CPU, 32GB RAM and RTX-2070S GPU.

D. Evaluation

We exploited the standard evaluation metrics for assessing classification performance, i.e. accuracy, weighted average precision, recall and F1, with the two labeled datasets. For the multi-label classification task, we calculated the exact match for accuracy and also reported Hamming loss, which is the fraction of labels that are incorrectly predicted. The evaluation results of the proposed S-BERT-KG and other models are shown in Table II and III and IV.

We report the precision, recall and F1 score of using the 4 BERT-based models for all seven different labels selected in our studies in Table II (for multi-class classification) and III (for multi-label classification). From both tables, we observe the proposed S-BERT-KG model significantly outperformed all other BERT-based models for labels: China, News, USA and Vaccine. For example, in the multi-class classification scenario (Table II), the S-BERT-KG model improved by around 21% F1 for tweets related to China, 20% F1 for tweets related to news, 19% F1 for tweets related to USA, 16% F1 for tweets related to vaccine, 13.30% in macro averaged F1 and 18.19% in weighted average F1. This confirmed that projecting S-BERT embedding to the knowledge graph embedding space could, in general, offer better performance for zero-shot classification.

One highlight from Table IV is that the BERT-CLS and the BERT-AVG models had the lowest performance in both classification scenarios. It shows that directly using the BERT model was not suitable for sentence classification without any training data. While applying the S-BERT model pre-trained with SNLI, MultiNLI and STS datasets, we observed significant improvement in terms of all evaluation metrics, e.g. around 10% in accuracy and 20% in F1 score in the multi-class scenario compared with BERT-CLS. This observation demonstrated the effectiveness of sentence representation with large S-BERT models pre-trained with Natural Language Inference (NLI) datasets. Table IV shows that the GloVe-AVG model outperformed the S-BERT with 12.49% in accuracy for multi-class classification and 14.17% in accuracy for multi-label classification, indicating that although S-BERT could learn good sentence-level embeddings, it might not generate semantically consistent word-level embeddings for labels.

Another notable observation in Table IV is that the proposed S-BERT-KG model significantly outperformed S-BERT-GloVe with respect to all the metrics considered. This indicated that the knowledge graph embedding could integrate common sense knowledge from the external knowledge base and improve model performance. When comparing S-BERT-KG with the recent BART-NLI model, we also observed significant improvement with 10.76% in accuracy, 10.76% in recall, 13.45% in F1 for multi-class classification, and 16.88% in recall, 16.82% in F1, 2.50% in hamming loss for multi-label classification. It should be noted that the BART-NLI model is based on a recent seq2seq architecture with a bidirectional encoder (e.g. BERT) and a left-to-right decoder (e.g. GPT), which outperformed BERT in NLI tasks.

In Table IV, we also reported the running time (t) of different models for multi-class classification and multi-label classification. The GloVe-AVG model that directly calculates the average value of word embedding vectors took the least time for prediction. If we compare BERT-CLS, S-BERT and BART-NLI, the increased time was related to the size of the models. For multi-class classification, the S-BERT model took 14.59 seconds while the S-BERT-KG model took 30.25 seconds. Thus, knowledge graph embedding alignment process consumed around 15 seconds to run. It is also noted that our proposed model shows a clear superiority to the BART NLI model in both efficiency and effectiveness. Since the S-BERT-KG model does not require a training process, which may take hours or days, it has the potential to be applied to real-time or near real-time streaming tweet classification.

For further comparison, we generated the t-SNE visualization[5] of the sentence and label embeddings (as shown in Figure 3) of the GloVe-AVG, S-BERT, and S-BERT-KG models. In the sub-figures, different colors represent different labels. By comparing 3a with 3b we can observe that the S-BERT model generates better sentence embeddings, i.e. the tweets related to China are better clustered. However, as the labels are poorly aligned in S-BERT, the overall performance was worse than GloVe-AVG as shown in Table IV. While using the S-BERT-KG, labels appeared much closer to their corresponding data clusters compared to the S-BERT model, and the sentences were also well clustered. With the visualization, we could further confirm that combing the sentence embedding model with the knowledge graph embedding is an effective method for leveraging unlabeled data for zero-shot tweet classification.

V. Conclusion and Future Work

Extracting useful information from enormous amount of social IoT data can be extremely challenging due to the lack of labeled quality data. Our study and experiments also confirmed that it is impractical to use the traditional supervised learning paradigm for deep neural network training. In addition, most deep learning models have not exploited the value of existing quality knowledge bases, usually in the form of graphs. Our current work manages to address these two issues and develops the S-BERT-KG model following the zero-shot learning paradigm for classification of COVID-19 related tweets. Performance of the S-BERT-KG model has been both impressive and promising, as evidenced by the evaluation results on both multi-class and multi-label classification tasks.

For future work, we plan to refine the proposed model in a number of directions. As we did not find more recent models pre-trained in the Sentence BERT architecture, we engaged the S-BERT model described in [17] for all the experiments and evaluation. It is expected that the S-BERT-KG model could be further improved with more recent models, e.g. roBERTa [24] and BART [25]. We plan to investigate the self-training method to further exploit the knowledge from the large amount

[5] https://lvdmaaten.github.io/tsne

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MODELS FOR ALL SEVEN LABELS IN MULTI-CLASS CLASSIFICATION IN TERMS OF PRECISION (P), RECALL (R), AND F1 SCORE (F1)

Labels	BERT-AVG			S-BERT			S-BERT-GloVe			S-BERT-KG		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Advice	45.53	40.88	43.08	41.67	47.45	**44.37**	36.21	15.33	21.54	44.05	27.01	33.48
China	63.84	25.17	36.10	92.59	55.68	69.54	91.88	55.46	69.17	91.26	90.65	**90.95**
Mask	100.0	04.89	09.32	73.64	78.22	**75.86**	79.52	58.67	67.52	74.75	67.11	70.73
News	21.57	97.73	35.34	56.82	08.09	14.16	67.14	15.21	24.80	68.93	45.95	**55.15**
Transportation	00.00	00.00	00.00	68.42	28.26	40.00	93.75	32.61	**48.39**	80.00	26.09	39.34
USA	100.0	05.67	10.74	82.65	17.02	28.22	68.63	29.41	41.18	68.65	53.36	**60.05**
Vaccine	00.00	00.00	00.00	09.87	93.75	17.86	09.76	96.88	17.73	20.72	89.58	**33.66**
macro avg	47.28	24.91	19.23	60.81	46.92	41.43	63.84	43.37	41.47	64.05	57.11	**54.77**
weighted avg	64.25	29.29	23.15	71.83	40.28	43.58	70.64	40.10	46.25	71.04	62.66	**64.44**

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT MODELS FOR ALL SEVEN LABELS IN MULTI-LABEL CLASSIFICATION IN TERMS OF PRECISION (P), RECALL (R), AND F1 SCORE (F1)

Labels	BERT-AVG			S-BERT			S-BERT-GloVe			S-BERT-KG		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Advice	23.38	41.94	30.02	34.23	32.90	**33.55**	23.33	09.03	13.02	25.00	05.81	09.42
China	50.38	24.19	32.68	95.54	61.91	75.14	90.62	62.82	74.20	95.59	90.07	**92.75**
Mask	82.86	10.66	18.89	83.63	69.49	**75.90**	89.51	47.06	61.69	96.46	40.07	56.62
News	25.84	58.58	35.86	49.28	08.33	14.26	42.72	10.78	17.22	60.00	31.62	**41.41**
Transportation	12.50	01.75	03.08	34.78	14.04	**20.00**	66.67	07.02	12.70	71.43	08.77	15.62
USA	62.79	13.59	22.34	75.51	18.62	29.88	68.35	25.00	36.61	64.40	60.40	**62.34**
Vaccine	14.29	01.74	03.10	10.86	91.30	19.41	11.65	96.52	20.79	18.84	87.83	**31.03**
macro avg	38.86	21.78	20.85	54.83	42.37	38.31	56.12	36.89	33.75	61.68	46.37	**44.17**
weighted avg	48.40	25.54	26.14	69.23	38.99	43.80	65.59	37.00	42.59	70.55	56.19	**58.76**

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT MODELS IN TERMS OF ACCURACY (A), PRECISION (P), RECALL (R), F1 SCORE (F1), HAMMING LOSS (H), AND RUNNING TIME (t) IN SECONDS

Methods	Multi-class Classification					Multi-label Classification					
	A	P	Rl	F1	t	A	P	R	F1	H	t
GloVe-AVG	52.76	65.15	52.76	46.93	**2.76**	32.05	75.96	40.10	42.73	12.93	**2.79**
BERT-CLS	30.96	50.26	30.96	23.83	10.84	3.76	24.46	37.92	28.76	32.73	11.42
BERT-AVG	29.29	64.25	29.29	23.15	10.85	10.20	48.40	25.54	26.14	19.95	11.24
S-BERT	40.27	71.83	40.28	43.58	14.59	17.88	69.23	38.99	43.80	17.77	15.64
S-BERT-GloVe	40.10	70.64	40.10	46.25	30.67	18.70	65.59	37.00	42.59	17.87	31.92
BART-NLI	51.21	**80.53**	51.21	50.41	107.19	**37.10**	**84.78**	39.31	41.94	14.17	116.75
S-BERT-KG	**62.66**	71.04	**62.66**	**64.44**	30.25	34.00	70.55	**56.19**	**58.76**	**12.67**	31.19

of unlabelled data, and exploit the few-shot learning technique when only a limited amount of labeled data is available. With the proposed zero-shot text classification architecture, we would like automatically create more labeled data for more comprehensive evaluation. Currently, all the labels used in this study are individual words. However, it may destroy its original semantics by representing key phrases with individual words using word embedding methods. We will further explore the performance of knowledge graphs in addressing this issue, and further apply the power of knowledge graphs, graph embeddings, and graph neural networks to other social IoT applications.

ACKNOWLEDGMENT

This research was funded by the Research Development Fund at Xi'an Jiaotong-Liverpool University, contract number RDF-16-01-34. It is also partially supported by the following: National Natural Science Foundation of China under no.61876155; Natural Science Foundation of Jiangsu Province BK20181189 and BE2020006-4B; Key Program Special Fund in XJTLU under no. KSF-T-06 and KSF-E-05.

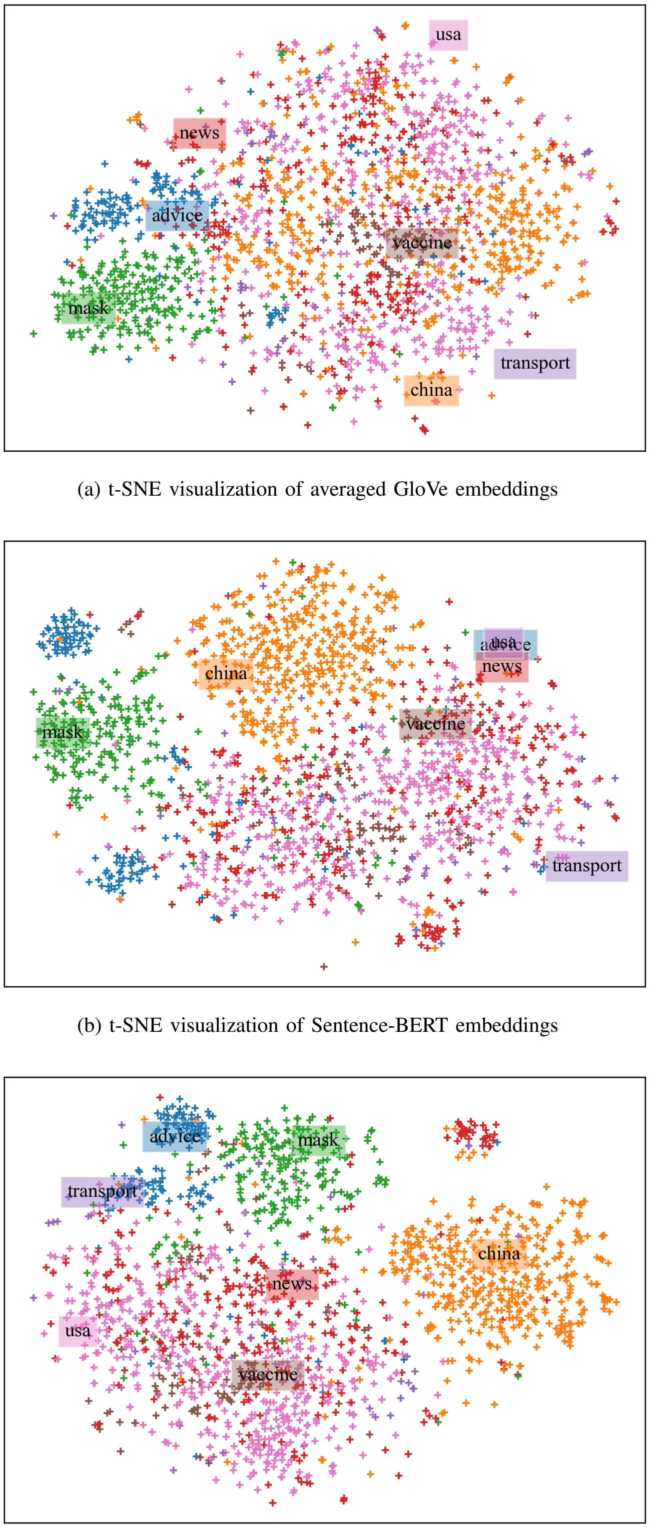
REFERENCES

[1] Amit Sheth. Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing*, 13(4):87–92, 2009.

[2] Antonio M Ortiz, Dina Hussein, Soochang Park, Son N Han, and Noel Crespi. The cluster between internet of things and social networks: Review and research challenges. *IEEE Internet of Things Journal*, 1(3):206–215, 2014.

[3] Jing Zeng, Laurence T Yang, Man Lin, Huansheng Ning, and Jianhua Ma. A survey: Cyber-physical-social systems and their system-level design methodology. *Future Generation Computer Systems*, 105:1028–1042, 2020.

[4] Puming Wang, Laurence T Yang, Jintao Li, Jinjun Chen, and Shangqing Hu. Data fusion in cyber-physical-social systems: State-of-the-art and perspectives. *Information Fusion*, 51:42–57, 2019.



(a) t-SNE visualization of averaged GloVe embeddings

(b) t-SNE visualization of Sentence-BERT embeddings

(c) t-SNE visualization of embeddings with Sentence-BERT to Knowledge Graph Projection

Fig. 3. t-SNE visualization of sentence and label embeddings of different methods

[5] Sina Dabiri and Kevin Heaslip. Developing a twitter-based traffic event detection model using deep learning architectures. *Expert Systems with Applications*, 118:425–439, 2019.

[6] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Eleventh International AAAI Conference on Web and Social Media*, pages 632–635, 2017.

[7] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*, 2016.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[12] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189, 2015.

[14] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, 2015.

[15] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Googles multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[16] Richong Zhang, Fanshuang Kong, Chenyue Wang, and Yongyi Mao. Embedding of hierarchically typed knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[17] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[18] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[19] Zhiwei Guo, Yu Shen, Ali Kashif Bashir, Muhammad Imran, Neeraj Kumar, Di Zhang, and Keping Yu. Robust spammer detection using collaborative neural network in internet of thing applications. *IEEE Internet of Things Journal*, 2020.

[20] Umair Qazi, Muhammad Imran, and Ferda Ofli. Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special*, 12(1):6–15, 2020.

[21] Catherine Ordun, Sanjay Purushotham, and Edward Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*, 2020.

[22] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*, 2019.

[23] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, 2019.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[26] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174, 2015.

[27] Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. Train once, test anywhere: Zero-shot learning for text classification. *arXiv preprint arXiv:1712.05972*, 2017.

[28] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*, 2019.

[29] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[30] Fan Zhou, Qing Yang, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Ashfaq Khokhar. Reinforced spatiotemporal attentive graph neural networks for traffic forecasting. *IEEE Internet of Things Journal*, 7(7):6414–6428, 2020.

[31] Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4883–4894, 2019.

[32] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

[33] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

[34] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[35] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1225–1234, 2016.

[36] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, 2019.

[37] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908, 2020.

[38] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, 2015.

[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[40] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[41] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[42] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[43] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[44] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

[45] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. 11 label propagation and quadratic criterion. 2006.