2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

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

Mr. P. Sreenivasa Rao, BALE SRIJA

Associate Professor, Department of Computer Science & Engineering.

M.Tech, Department of CSE, (21JJ1D5801), balesrija@gmail.com

ABSTRACT: 'Citizen sensing' and 'human as monitors' are pivotal ideas for cyber-physical-social systems' social Internet of Things. It's easy to get social media data from the social world. This data has become useful for study in many fields, such as evaluating crises and disasters, finding social events, and the new COVID-19 analysis. It would be better for everyone if there were faster and more accurate ways to process and study useful information, like knowledge gathered from social data. Deep neural network improvements have made a big difference in how well many social media research jobs work. DL models, on the other hand, need a lot of labeled data to train, but most CPSS data isn't labeled, so standard methods can't be used to make good learning models. Also, the most advanced Natural Language Processing (NLP) models that have already been taught don't use knowledge graphs, which means they often don't work well in real-world situations. We come up with a new zero-shot learning method that solves the problems by making good use of current knowledge graphs to sort through a huge amount of social text data. The suggested system was tried on a few genuine Coronavirus tweets. Using the latest DL models for natural language processing, it outperforms six traditional models.

Keywords *–Internet of Things (IoT), knowledge graph, natural language processing (NLP), social media data analysis, zero-shot learning.*

1. INTRODUCTION

The concept of humans acting as sensors or citizen sensing is gaining popularity as smart devices, the IoT, mobile social networks, and cloud computing become more widespread. In this term, people are both the customers and providers of data. Everyone can use it to gather, examine, report, and share knowledge, which helps them see and understand the world better. At the same time, it is very important for the growth of social IoT, which is a key part of Cyber-Physical-Social systems (CPSS). A huge amount of information from social media can be gathered and then used in different jobs that could have a big effect on society as a whole. For instance, Twitter users can post real-time

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

traffic information, which makes it easier to find traffic events. Other examples are accounts of people who are hurt or missing, damage to infrastructure, and warnings and cautions. All of these help with assessing the crisis or disaster and responding to it. Usually, Natural Language Processing (NLP) methods are used to get useful data and information from social media. Deep Neural Networks (DNNs) have recently excelled in tasks like natural language processing and image processing. In the supervised learning model, DNNs are currently the most effective classifiers, as long as there is a substantial amount of accurately labeled data available. Application fields include car recognition from photos, document grouping, and neural machine translation. Lack of labeled data frequently causes them to fail. You can address this issue by applying what you learnt from solving one problem to a related one. We call this transfer learning. A new method in natural language processing (NLP) involves using transfer learning by training models on a large collection of unlabeled text and then using those trained models for a specific task.

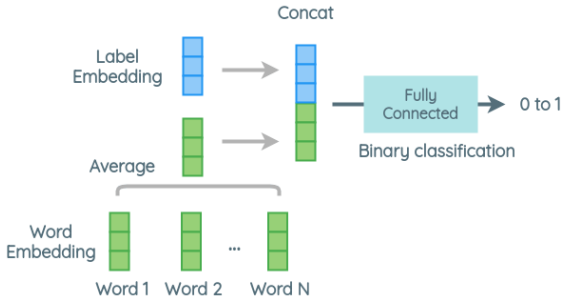

Fig.1: Example figure

Pre-trained models include word2vec[8], GloVe[9], and BERT[10]. The models have been used to caption images[11], analyze social media sentiment[12], and categorize text in smart city applications[6]. Beyond employing these models, researchers are researching domain adaptation[13], multitask learning[14], and zero-shot learning[15] for transfer learning. A classifier must identify instances from classes not observed during training for zero-shot learning to operate. It's ideal for managing and researching social media data since much of it is unlabeled and hard to classify.

2. LITERATURE REVIEW

The cluster between Internet of Things and social networks: Review and research challenges:

The IoT-SN cluster connects individuals to the ubiquitous computing cosmos. The IoT provides environmental data, while the SN enables human-to-device interactions. Social Internet of Things (SIoT) is a new paradigm for pervasive computing beyond IoT. Early research on social-driven IoT only leverage SIoT features to enhance a few performance factors. This article begins with a comprehensive understanding of SIoT and critical views to picture actual ubiquitous computing. This paper starts with a literature review and discusses the advancement of IoT research from Intranet of Things to SIoT. It then presents a SIoT architecture and

discusses its potential for innovation, research challenges, and unresolved issues.

A survey: Cyberphysical-social systems and their system-level design methodology:

Cyber-physical-social systems (CPSS) fundamentally change how people, computers, and the physical world interact. This study examines the development of CPSS via CPS, CSS, and similar methods. After conducting a review of existing literature, the research on the Internet of Things (IoT) has evolved from focusing on internal networks (Intranet) to a more advanced concept called Social Internet of Things (SIoT). This study suggests a SIoT architecture that can facilitate further advancements in the field, while also highlighting the research challenges and unanswered questions that still remain.

Data fusion in cyberphysical-social systems: State-of-the-art and perspectives:

CPSSs complement CPSs by smoothly integrating cyber, physical, and social spaces. CPSSs spread information from single to tri-space to revolutionize data science. A complete overview of data fusion in CPSSs is presented in this study. To explain information fusion in CPSS, we first study data collection and representation and suggest using tensors to describe CPSS data. Then, we define CPSS data fusion broadly. Then, sample CPSS data fusion strategies are discussed. We also present CPSS data fusion algorithms using tensors. We also examine data fusion framework design and present a complete CPSS framework. Challenges and future work are also highlighted.

Developing a twitter-based traffic event detection model using deep learning architectures:

Researchers have been using Twitter data to monitor traffic incidents. They convert tweets into numerical feature vectors using a method called bag-of-words, which ignores the order of words in a tweet and has some concerns regarding size and sparsity. To address these concerns, researchers use traffic-related keywords to reduce the dimensionality of the bag-of-words. However, there are concerns that these pre-defined keywords may not cover all traffic-related terms and that the language used in tweets changes over time. To overcome these issues, deep-learning models are used to encode tweets as numerical vectors and classify them into different categories related to traffic. This process involves converting words into low-dimensional vectors and analyzing their meaning. Traffic events are identified using CNN and RNN word-embedding algorithms. Our model is trained and tested using a huge amount of traffic tweets received via Twitter API endpoints and tagged efficiently. Trials on our labeled dataset reveal that the suggested strategy outperforms current approaches.

Robust classification of crisis-related data on social networks using convolutional neural networks:

Social media, particularly Twitter, is becoming recognized as a source of actionable and tactical information during catastrophes. ML approaches, especially supervised learning, struggle to analyze enormous social media crisis data in real time. Labeled data shortages hinder ML early in a crisis. For optimal training and feature creation, modern classification methods need a lot of labeled event-specific data. We provide binary and multi-class neural network-based tweet categorization techniques. Our neural network-based models beat state-of-the-art methods without feature engineering. We suggest using data from outside of the disaster event to achieve positive outcomes in the initial stages of a crisis, even without data that has been labeled or categorized.

3. METHODOLOGY

Recently, there has been a lot of interest in studying how to use current, good knowledge sources successfully in DNNs. Many knowledge bases and knowledge graphs already exist, and the information they store is made up of facts and human knowledge that has been gathered over hundreds of years. Putting this kind of information into learning tools could be very helpful. Systems needn't learn everything from start. However, categorization errors are mostly avoidable. As a method for prediction, reasoning, data mining, and finding information, embedding has become very important. More and more research is being done on graph embedding methods that use vectors to show the systematic organization of a knowledge base. If learning systems get rich structural information from knowledge sources, prediction, classification, and recommendation functions should improve. Complex DL and knowledge graph embedding research issues have not been completely investigated.

Disadvantages:

1. DL models, on the other hand, need a lot of labeled data to be trained, but most CPSS data isn't labeled, so standard methods can't be used to make good learning models.
2. Hasn't been studied in depth.

This paper proposes a new method for categorizing social text data, specifically COVID-19 tweets, without the need for training data. The method utilizes existing knowledge graphs to achieve this. Utilizing a pre-trained sentence-based BERT model, tweets are embedded in a category-matching space. Because of class names, this inserting might be less steady than word-level embeddings. To address this, a comprehensive knowledge graph called ConceptNet is used to create a model that represents the labels. Linear projection links the

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

knowledge network to sentence embedding. The suggested model, S-BERT-KG, sorts COVID-19 tweets without tag training.

Advantages:

1. The S-BERT-KG model has been proven to be superior to other standard models.
2. Make predictions that are pretty accurate without using any labeled data.

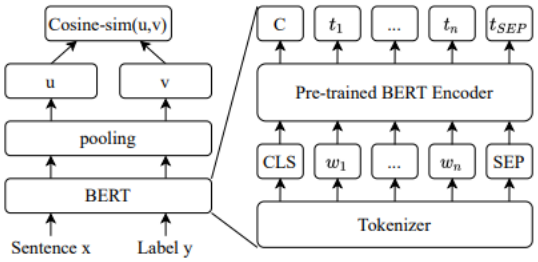

Fig.2: System architecture

MODULES:

To carry out the project mentioned earlier, we have created the following modules.

- Data exploration: Data will be imported using this module.
- Processing: This module reads and processes data.
- Splitting data into train & test: This module will separate data into train and test.
- Model generation: Building the model - GCN - GCN with BERT - GRU - LSTM - CNN - Bi- LSTM - BERT GCN + LSTM + CNN (Zero-Short Model) and Ensemble CNN+LSTM. Algorithms accuracy calculated.
- User signup & login: You may register and log in using this module.
- User input: This module aids forecasting.
- Prediction: Last forecast

4. IMPLEMENTATION

ALGORITHMS:

GCN: Graph Convolutional Networks (GCNs) provide semi-supervised learning on graph-structured data. It uses efficient graph-based convolutional neural networks. This shares weights in each recurrent step like an RNN. Grec below has the same settings, but GCN does not share weights across hidden levels.

GCN with BERT: BERT is an open-source NLP ML framework. To assist computers interpret ambiguous material, BERT uses surrounding text to provide context. BERT, a deep bidirectional language representation trained on plain text, is a pre-trained model that H2O.ai uses to achieve advanced natural language processing.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

GRU: Kyunghyun Cho et al. introduced GRUs in 2014 for recurrent neural networks. Due to its absence of an output gate, the GRU has fewer parameters than an LSTM with a forget gate. Learn how GRU works. We have a GRU cell that resembles an LSTM or RNN cell. At each timetamp t, it receives Xt and Ht-1 from t-1. New hidden state Ht is output and sent to the next timestamp.

LSTM: LSTMs, which are a type of RNN, are responsible for enabling DL by learning long-term dependencies and order dependence for sequence prediction. Machine translation, speech recognition, and other complex issues need this behavior. LSTMs in DL are difficult.

CNN: CNNs are DL network architectures used for image recognition and pixel data processing. CNNs are the preferred DL neural network design for object recognition. CNN uses convolution layers, pooling layers, and fully linked layers to automatically and adaptively learn feature spatial hierarchies via backpropagation.

Bi-LSTM: A BiLSTM layer develops bidirectional long-term associations between time series or sequence data stages. The network may learn from the whole time series at each step with these dependencies. It learns sentence context from BiLSTMs by recognizing what words follow and precede a word.

BERT GCN + LSTM + CNN (Zero-Short Model): Zero-Shot Learning evaluates test data from untrained classes using a pre-trained model. The model must expand to new categories without semantic information. Retraining models are unnecessary with such learning frameworks.

Ensemble CNN+LSTM: Ensemble modeling uses numerous modeling algorithms or training data sets to predict an outcome. Ensemble models combine base model predictions to make one final forecast for unknown data.

5. EXPERIMENTAL RESULTS

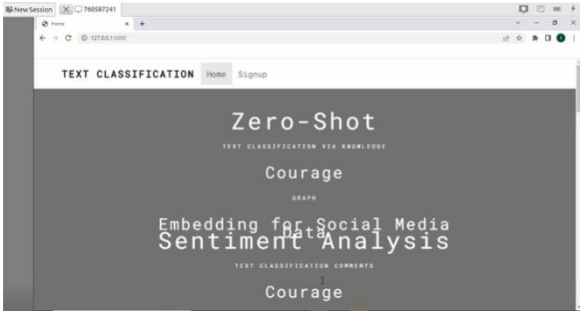

Fig.4: Home screen

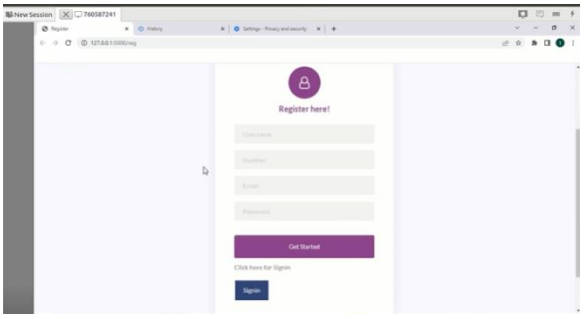

Fig.5: User registration

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

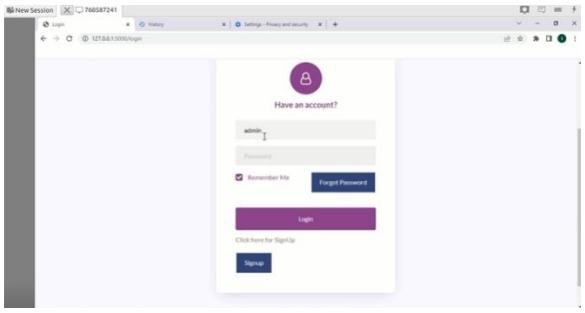

Fig.6: user login

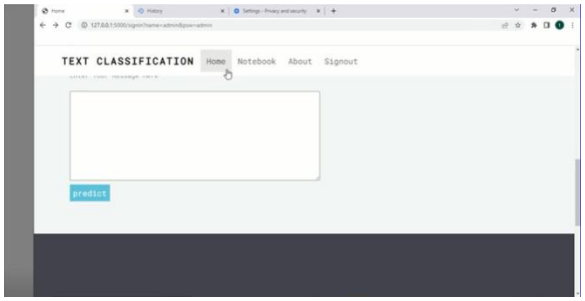

Fig.7: Main screen

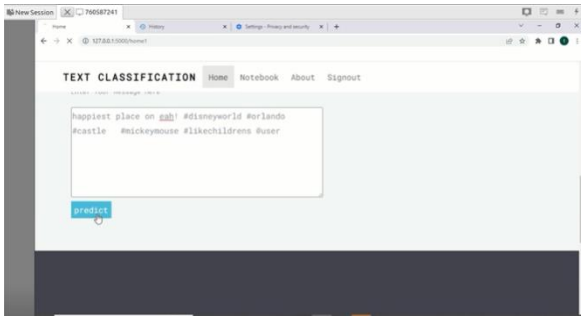

Fig.8: User input

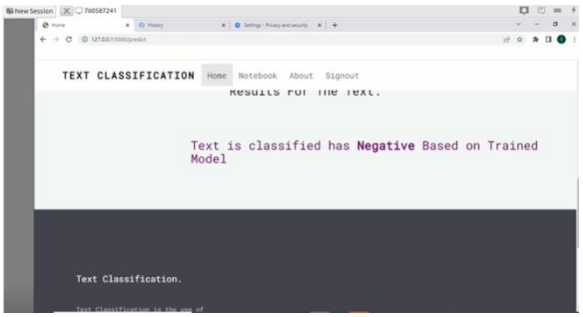

Fig.9: Prediction result

6. CONCLUSION

The massive social IoT data without annotated quality data makes it difficult to find meaningful information. Our study and testing showed that the typical supervised learning approach can't train DNNs. Graphs are excellent information sources, yet most DL models don't utilize them. This work addresses these two issues by creating the S-BERT-KG model to categorize COVID-19 tweets using zero-shot learning. The S-BERT-KG model performed well on multiclass and multilabel classification tests and showed potential. We wish to improve the intended model for future work in many ways. We couldn't discover any newer models trained in the S-BERT design, so we utilized it for all testing and assessments. Newer models like roBERTa and BART may improve S-BERT-KG. To maximize knowledge from large volumes of unlabeled data, we want to employ self-training, and when we have little labeled data, we want to use few-shot learning. The goal is to improve the zero-shot text categorization algorithm by generating more labeled data. Currently, all names used in this algorithm are single words, but this may lead to important phrases being replaced by single words in word embedding. We will explore the use of knowledge graphs in addressing this challenge and in other social IoT applications.

REFERENCES

[1] A. Sheth, "Citizen sensing, social signals, and enriching human experience," IEEE Internet Comput., vol. 13, no. 4, pp. 87–92, Jul. 2009.

[2] A. M. Ortiz, D. Hussein, S. Park, S. N. Han, and N. Crespi, "The cluster between Internet of Things and social networks: Review and research challenges," IEEE Internet Things J., vol. 1, no. 3, pp. 206–215, Jun. 2014.

[3] J. Zeng, L. T. Yang, M. Lin, H. Ning, and J. Ma, "A survey: Cyberphysical-social systems and their system-level design methodology," Future Gener. Comput. Syst., vol. 105, pp. 1028–1042, Apr. 2020.

[4] P. Wang, L. T. Yang, J. Li, J. Chen, and S. Hu, "Data fusion in cyberphysical-social systems: State-of-the-art and perspectives," Inf. Fusion, vol. 51, pp. 42–57, Nov. 2019.

[5] S. Dabiri and K. Heaslip, "Developing a twitter-based traffic event detection model using deep learning architectures," Expert Syst. Appl., 118, pp. 425–439, Mar. 2019.

[6] D. T. Nguyen, K. Al-Mannai, S. R. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks," in Proc. 11th Int. AAAI Conf. Web Soc. Media, 2017, pp. 632–635.

[7] M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: Humanannotated twitter corpora for NLP of crisis-related messages," 2016. [Online]. Available: arXiv:1605.05894.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: arXiv:1301.3781.

[9] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), 2014, pp. 1532–1543.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: arXiv:1810.04805.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, 2015, pp. 3156–3164.

[12] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdiscipl. Rev. Data Min. Knowl. Discov., vol. 8, no. 4, p. e1253, 2018.

[13] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in Proc. Int. Conf. Mach. Learn., 2015, pp. 1180–1189.

[14] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-task learning for multiple language translation," in Proc. 53rd Annu. Meeting Assoc. Comput. Linguist. 7th Int. Joint Conf. Nat. Lang. Process. (Volume 1: Long Papers), 2015, pp. 1723–1732.

[15] M. Johnson et al., "Google's multilingual neural machine translation system: Enabling zero-shot translation," Trans. Assoc. Comput. Linguist., vol. 5, no. 2, pp. 339–351, Oct. 2017.