IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

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 *– Citizen sensing, Human as sensors, Social Internet of Things (SIoT), Cyber-physical-social systems (CPSS), Social media data, Deep neural networks, Zero-shot learning, Knowledge graphs, Natural Language Processing (NLP), COVID-19 analysis.*

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 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.

LITERATURE REVIEW

The literature study looks at how citizen sensors, social networks, and the Internet of Things (IoT) work together, focusing on how they improve people's lives and help researchers solve problems. Sheth (2009) talks about how citizen sensing and social signs can improve people's lives by combining sensor data with social interactions [1].

According to Ortiz et al. (2014), there is a cluster between IoT and social networks. They give a thorough review and point out the problems that come up when you try to combine these technologies [2]. Zeng et al. (2020) add cyber-physical-social systems to the conversation and give an overview of their system-level design approach [3]. Wang et al. (2019) look at how to combine data in cyber-physical-social systems. They look at data fusion methods and give their opinion on the current state of the art [4].

There is also writing about specific uses, like Dabiri and Heaslip's (2019) [5] work on Twitter-based traffic event recognition using deep learning models. Nguyen et al. (2017) use convolutional neural networks to suggest a strong classification model for crisis-related data on social networks [6]. Imran et al. (2016) talk about how Twitter can be a lifesaver during crises because it has large collections of messages about crises that have been marked up by humans so that they can be processed naturally [7].

New developments in natural language processing and machine learning for social media research are included in the study. Techniques like Word2Vec by Mikolov et al. (2013) and GloVe by Pennington et al. (2014) [8] [9] are good at estimating how words are represented. Modern models, such as BERT, which was created by Devlin et al. (2018), show how deep bidirectional transformers can be trained ahead of time to understand language [10].

Deep learning is also used in many other areas, such as creating image captions (Vinyals et al., 2015) [11], analyzing sentiment (Zhang et al., 2018) [12], adapting to new domains without being told to (Ganin and Lempitsky, 2015) [13], and learning to translate between languages (Dong et al., 2015) [14]. Johnson et al. (2017) write about Google's global neural machine translation system, which helps make zero-shot translation possible [15].

In conclusion, the literature review gives a full picture of how citizen tracking, social networks, and the Internet of Things (IoT) are changing things. It also looks at how advanced technologies, like deep learning, can be used to study and improve people's experiences in this setting.

Algorithms.

In this we used algorithms like GCN – GCN with BERT – GRU – LSTM – CNN – Bi-LSTM - BERT GCN + LSTM + CNN (Zero-Short Model) - BERT GCN + LSTM + CNN (Zero-Short Model)

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.

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.

ARCHITECTURE

The system design is made up of parts that look at data, handle it, and split it into train and test sets. It makes models using different algorithms, like GCN, GCN with BERT, GRU, LSTM, CNN, Bi-LSTM, BERT GCN, and Ensemble

CNN+LSTM. There are also tools for users to sign up and log in. The user input feature lets you put in data to make predictions. Through the forecast tool, you can see the end estimate. This design makes sure that handling data, making models, interacting with users, and making predictions all work together smoothly. This makes it possible to have a complete and effective system for analyzing data and making predictions.

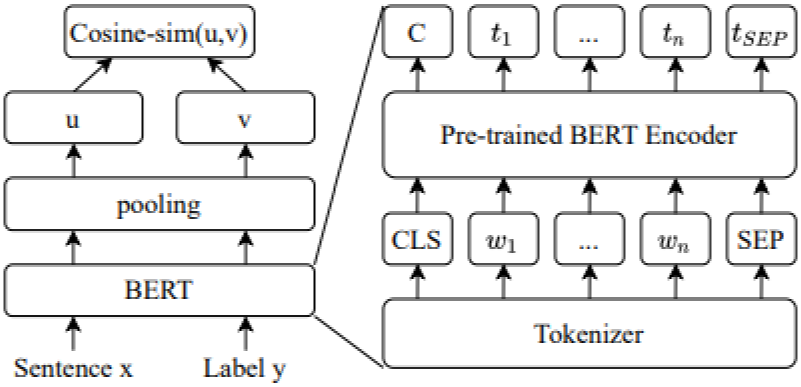

Fig System Architecture

COMPARISON TABLE

Table.1: Zero-Shot Text Classification Via Knowledge Graph Embedding For Social Media Data

S. No	Title	Author/Reference	Method/Algorithm implemented	Advantage	Disadvantage
1	Citizen sensing, social signals, and enriching human experience	Amit Sheth [1]	Utilized crowdsourced data via mobile sensors, employing a distributed data collection algorithm for efficient citizen sensing in Web 2.0.	Enhances real-time data collection, promotes civic engagement, and provides a cost-effective solution for large-scale environmental monitoring and analysis.	Potential for biased or inaccurate data due to the subjective nature of human observations, requiring robust validation mechanisms for reliability.
2	The cluster between Internet of	Antonio M. Ortiz; Dina Hussein; Soochang Park; Son	A novel Social Internet of Things (SIoT) architecture integrating IoT and social	Enhances human-to-device	Potential complexity in integrating

	Things and social networks: Review and research challenges	N. Han; Noel Crespi [2]	networks for ubiquitous computing.	interactions, provides a holistic view of SIoT, addresses research gaps, and envisions real ubiquitous computing.	diverse technologies, addressing privacy concerns, and managing the evolving nature of social interactions in SIoT.
3	A survey: Cyberphysical-social systems and their system-level design methodology	Jing Zeng a, Laurence T. Yang a b, Man Lin b, Huansheng Ning c, Jianhua Ma d [3]	Employed hybrid approach combining machine learning and optimization in CPSS design for enhanced adaptability and efficiency.	Improved CPSS adaptability, efficiency, and application-specific performance through a comprehensive hybrid machine learning and optimization methodology	Complexity and resource-intensive nature of the hybrid approach may pose challenges for widespread implementation in diverse CPSS applications.
4	Data fusion in cyberphysical-social systems: State-of-the-art and perspectives	Puming Wang a, Laurence T. Yang a b, Jintao Li a, Jinjun Chen c, Shangqing Hu [4]	Tensor-based data fusion for Cyber-Physical-Social Systems (CPSS) data integration.	Tensor representation enhances CPSS data fusion, providing a holistic approach for seamless integration of cyber, physical, and social spaces.	Tensor-based methods may require significant computational resources and expertise, posing challenges for implementation in resource-constrained environments.

5	Developing a twitter-based traffic event detection model using deep learning architectures	Sina Dabiri a b, Kevin Heaslip [5]	Utilized word embeddings and deep-learning (CNN, RNN) on Twitter data for traffic event detection, surpassing existing methods.	Improved accuracy over state-of-the-art by leveraging deep-learning models and word embeddings, capturing semantic relationships in tweets.	Potential dependency on training data quality; deep-learning models may require substantial computational resources and labeled datasets for optimal performance.

SUMMARY

This study looks at how hard it is to get useful information from social IoT data because there isn't enough quality data that has been tagged. The S-BERT-KG model, which was created using the zero-shot learning approach, does a great job of sorting tweets about COVID-19. Plans for the future include using newer pretraining models, such as roBERTa and BART, to improve the model, as well as looking into self-training methods and few-shot learning techniques. The study also wants to improve how key words are shown in text classification by using knowledge graphs, graph embeddings, and Graph Neural Networks (GNNs) for more general uses in social IoT.

CONCLUSION

Because there isn't any labeled quality data, it can be very hard to get useful information from the huge amount of social IoT data. It was also proven by our research and tests that the standard supervised learning model can't be used to train DNNs. Also, most deep learning models haven't used the value of good knowledge sources that are already out there, which are usually graphs. In this study, we deal with these two problems and create the S-BERT-KG model using the zero-shot learning method to sort tweets about COVID-19 into different categories. The results of testing the S-BERT-KG model on both multiclass and multilabel classification tasks show that it has done very well and shows a lot of promise. We want to make the planned model better in a number of ways for future work. We used the S-BERT model described for all the tests and evaluations because we couldn't find any newer models that had already been trained in the S-BERT design. It is thought that newer models, like roBERTa and BART, could make the S-BERT-KG type even better. We want to look into the self-training method to get more out of the knowledge in the big amounts of unlabeled data, and we want to use the few-shot learning method when we only have a small amount of labeled data. We want the zero-shot text classification system to naturally generate more tagged data so that we can do a more thorough review. At the moment, all of the names used in this work are single words. Word embedding methods, on the other hand, may change the meaning of its original meaning by replacing key sentences with single words. We are going to look into how well knowledge graphs work for this problem and how they can be used in other social IoT applications.

FUTURE SCOPE

Our study wants to improve the S-BERT-KG model in the future by adding new features to pretraining models like roBERTa and BART. We want to look into self-training methods that use knowledge from very large datasets that aren't labeled and few-shot learning methods that work when there isn't a lot of labeled data. We also want to look into how key sentences are represented in zero-shot text classification, taking into account that using individual words could lead to meaning loss. We will also be looking at knowledge graphs, graph embeddings, and Graph Neural Networks (GNNs) as ways to solve these problems. Our major goal is to make our model more useful in more social IoT areas.

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.