ISSN: 2641-2039

Global Journal of
Engineering Sciences

DOI: 10.33552/GJES.2021.08.000698

Iris Publishers

Mini Review

Opportunities and Challenges of Deep Learning Models for Short Text Data Analysis

Ashis Kumar Chanda*

Adjunct faculty of Rowan University, Glassboro, New jersey, USA

*Corresponding author: Ashis Kumar Chanda, Adjunct faculty of Rowan University, Glassboro, New jersey, USA.

Received Date: November 08, 2021

Published Date: November 19, 2021

Abstract

The rapid growth of the internet usage on mobile devices has increased the opportunities to communicate with other people and publicly share news and opinions. As a result, a huge amount of data is generated every day, and it becomes a great resource for investigating different research problems such as sentiment analysis, entity finding, and prediction problems. Recent advanced machine learning methods use deep neural network models to solve the problems that require huge datasets to train and converge the models. However, people mostly use short text data for such online communication purposes. Analyzing the sentiment of the short text is challenging because of its natural characters, such as sparseness, immediacy, and misspelling. Therefore, this study focuses on the opportunities and challenges of deep learning models for short text data analysis. The paper discussed research works that applied deep learning (deep neural network) models to analyze short text data and explained unsolved problems in this area.

Keywords: Short text data; natural language processing; NLP; deep neural network models; social media data; Twitter data

Introduction

People tend to share their experiences, emotions, and news with other people using different online social media platforms (i.e., Twitter, Facebook), where they mostly use short texts. The short text has also been used in many other fields, such as mobile short messages, news titles, and blog comments. Although there is no fixed length limit to define a text as short, usually, the text length remains smaller than 200 characters. For example, a tweet can contain up to 280 characters on Twitter.

Since the online social media data is publicly available, researchers became interested in working on the data to solve several research problems such as knowledge extraction [1], text generation [2], and natural disaster prediction [3]. However, it is challenging to analyze short text data because of their shortness, sparsity (i.e., diverse word content) [4], velocity (rapid growth of short text like SMS and tweet), and misspelling [5]. For example, a short text containing only a few words does not provide enough shared context for a good similarity measure. Moreover, the short texts are sent im mediately in real-time, and the description is concise. People often use several abbreviations, local or non-standard terms, and noise or symbols in their real-time short text communication. Therefore, analyzing short text data has become a daunting research topic in natural language processing (NLP) fields.

Related works and Challenges

Traditional methods used rule-based models and learning-based models to analyze short text for different purposes but faced difficulty because of the characteristics of short text. For example, the authors of [6] discussed the difficulty of traditional methods on short text classification tasks. Another research work [2] focused on the automatic generation of short-text conversation using a retrieval-based automatic response model. Existing research focused on linking short text data such as tweets to news [7], where they used a graph-based latent variable model to find the inter short text correlations. They also showed that tweet-specific features (hashtag) and news specific features (named entities), and temporal constraints could extract text-to-text correlations, thus completing the semantic picture of a short text. Moreover, a recent survey paper [8] investigates how traditional methods work on topic modeling problems when applied to short textual social data. These methods are latent semantic analysis, latent Dirichlet allocation, non-negative matrix factorization, random projection, and principal component analysis.

Recent advances in machine learning models show that deep learning models can solve several complex problems for large datasets. However, it is also interesting to discover how the deep learning models work for short-length text data. For example, Jiaming et al. [9] applied Convolutional neural networks (CNN) to cluster short text on social media data, and they found that the deep learning method overcame traditional methods. Another research work performed on hate speech classification [10] also used deep learning methods on short text data. However, deep learning models require huge labeled data to train models. Thus, labeling short text data is another potential challenge. For this reason, Linmei et al. [11] proposed a semi-supervised model (Heterogeneous Graph ATtention networks (HGAT)), leveraging the full advantage of a few labeled data and large unlabeled data through information propagation along with the graph. The authors showed that the attention mechanism could learn the importance of different neighboring nodes (words) and the importance of different node (information) types to a current node.

A new language representation model, BERT (Bidirectional Encoder Representations from Transformers), is [12] is proposed recently that constructs different vectors for the same word in different contexts. In the past, a single low-dimensional vector representation or embedding was used to present a word from a given document. BERT embeddings have been successfully used in several natural language processing, including short text analysis tasks. The research work of [3] showed that deep learning models (i.e., BiLSTM) with the BERT embedding have the best accuracy in predicting disaster-type tweets from Twitter data where the mean length of the tweet is 12.5 words. Another work [13] showed that an advanced deep neural network (i.e., BiLSTM-CNN with attention) could improve disaster-type tweet prediction tasks.

Although deep learning methods have been successfully used for short text analysis tasks, there are still many open challenges. For example, previous research works removed symbols or emojis from a short text at the time of data processing steps. However, it is essential to explore how the symbols or emojis have an effect on the sentiment analysis task. Marco, et al. [14] provided an extensive research study to explain the data pre-processing steps, which plays a vital role in short text data analysis. Moreover, qualitative study is also needed to analyze the performance of the deep learning models on short text data to observe when and how the models work successfully [15].

Conclusion

The number of works on deep learning for short text data analysis has grown explosively in recent years. Such works have achieved accuracy comparable to that of traditional feature-based approaches. Specifically, we found that deep learning models with contextual embeddings (BERT) yield the best results. However, some new challenges and problems related to interpretability, scalability, and efficiency must be addressed. Furthermore, it is also worth investigating new applications from the perspectives of datasets and methods.

Acknowledgement

None

Conflict of Interest

None.

References

1. Mai Monica, Carson KLeung, Justin MC Choi, Long Kei, Ronnie Kwan(2020) Big data analytics of Twitter data and its application for physician assistants : who is talking about your profession in Twitter?." In Data Management and Analysis, pp. 17-32. Springer, Cham.
2. Wang Hao, Zhengdong Lu, Hang Li, Enhong Chen (2013) A dataset for research on short-text conversations. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 935-945.
3. Chanda, Ashis Kumar (2021) Efficacy of BERT embeddings on predicting disaster from Twitter data. arXiv preprint arXiv:2108.10698
4. Chen Mengen, Xiaoming Jin, Dou Shen (2011) Short text classification improved by learning multi-granularity topics. In the Twenty-second international joint conference on artificial intelligence.
5. Issa Alsmadi, Keng Hoon Gan (2019) Review of short-text classification. International Journal of Web Information Systems 15(2): 155-182.
6. Song Ge, Yunming Ye, Xiaolin Du, Xiaohui Huang, Shifu Bie (2014) Short text classification: A survey. Journal of multimedia 9(5): 635.
7. Guo, Weiwei, Hao Li, Heng Ji, Mona Diab (2013) Linking tweets to news: A framework to enrich short text data in social media. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 239-249.
8. Albalawi Rania, Tet Hin Yeap, Morad Benyoucef (2020) Using topic modeling methods for short-text data: A comparative analysis. Frontiers in Artificial Intelligence 3: 42
9. Xu Jiaming et al. (2015) Short text clustering via convolutional neural networks. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pp. 62-69.
10. Rizos Georgios, Konstantin Hemker, Bjorn Schuller(2019) Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 991-1000.
11. Linmei Hu, Tianchi Yang, Chuan Shi, Houye Ji, Xiaoli Li (2019) Heterogeneous graph attention networks for semi-supervised short text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4821-4830.
12. Devlin Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

13. Song Guizhe, Degen Huang (2021) A Sentiment-Aware Contextual Model for Real-Time Disaster Prediction Using Twitter Data. Future Internet 13(7): 163.

14. Pota Marco, Mirko Ventura, Hamido Fujita, Massimo Esposito (2021) Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. Expert Systems with Applications 181: 115119.

15. Guo, Weiwei, Hao Li, Heng Ji, Mona Diab (2013) Linking tweets to news: A framework to enrich short text data in social media. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 239-249.

Citation: Ashis Kumar Chanda. Opportunities and Challenges of Deep Learning Models for Short Text Data Analysis. Glob J Eng Sci. 8(5): 2021. GJES.MS.ID.000698. DOI: 10.33552/GJES.2021.08.000698.