

Towards Improving Text Classification Tasks Based on Knowledge Graphs for Limited Labeled Data

Hongzhi Zhang^{†,*}, M. Omair Shafiq[†]

[†]School of Information Technology, Carleton University, Ottawa, ON, Canada

Abstract

Pre-trained Transformer models have become popular in various Natural Language Processing (NLP) tasks, following a two-step process of 'pre-training' and 'fine-tuning'. However, with the abundance of information on the web, some domain knowledge may be lacking. This can result in poor performance during the fine-tuning step when there is limited training data available. To address this issue in the case of limited data, we propose a knowledge graph-based data expansion method that enables the model to achieve good results even when there is limited data in the fine-tuning step. We extract entities in the text through Named Entity Recognition and then search for related information in the knowledge graph to expand the text's content. This allows the pre-trained model to acquire more external knowledge and enhance its training. We used our data expansion method to conduct experiments on the following well-known models, i.e., BERT, RoBERTa, and GPT-3. Our experiments show that our approach can improve the accuracy of language models on text classification tasks when training data is limited.

Keywords: Natural Language Processing, Text Classification, Data Expansion, Language Models, Knowledge Graph

1. Introduction and Background

Since the introduction of Transformer models [1] by Google in 2017, many variants of the Transformer model, such as BERT [2] and RoBERTa [3], have become popular. These pre-trained models use large corpora and have shown great success in natural language processing tasks. However, when it comes to fine-tuning for a specific task, these models still require a significant amount of data for optimal performance. This can pose a challenge when there is a limited amount of data available for a certain domain. To address this issue, we propose using the well-known knowledge graphs (KG) as a means of data expansion. Our approach involves identifying entities in the text through named entity recognition and then retrieving related information from the knowledge graph to expand the information contained in the text. This enhances the model's ability to obtain external knowledge and leads to improved performance, even with limited training data. Our results show that our proposed method can effectively improve the performance of the model with small amounts of training data.

In addition to our previous experiments, we also conducted experiments with GPT-3 [4] and obtained similar positive results. This demonstrates the effectiveness of incorporating knowledge graphs as a means of data expansion and also provides a prompt method for introducing knowledge graphs.

1.1. Text classification tasks

Text classification is an essential task in Natural Language Processing (NLP), whereby text data is categorized based on predefined classification criteria. With the growing amount of text data available online, manual classification is both time-consuming and prone to errors due to human subjectivity. As a result, the use of deep learning methods for automated

*hongzhizhang@email.carleton.ca

text classification is essential in overcoming these limitations. By automating the repetitive and tedious task of text annotation, time is saved, and the resulting annotations are consistent and of high quality. In the domain of product development, sentiment analysis is a common NLP task used to evaluate customer feedback on e-commerce platforms. This analysis allows businesses to identify the strengths and weaknesses of their products from different perspectives and make improvements accordingly. Short text classification poses a challenge due to the limited amount of information available.

To assess the effectiveness of our proposed approach in improving model performance with limited training data, we conducted experiments on two challenging tasks: sarcasm detection and news headline classification. The former is a binary classification problem that presents a significant challenge due to the difficulty humans face in distinguishing between sarcastic and sincere comments. The latter is a multi-class classification problem with 25 discrete categories, where the data distribution is imbalanced, with some categories having considerably more instances than others. These differences in data availability create a significant obstacle to achieving high classification accuracy.

1.2. Knowledge graph

The components of a knowledge graph are entities, relationships, and attributes. These elements collectively describe the connections and information between various objects and concepts:

- **Entities:** Refers to things that exist objectively and can be distinguished from each other. It can be a concrete person, location, or thing, or it can be an abstract concept or connection. Entities are the most basic elements in knowledge graphs.
- **Relationships:** In the knowledge graph, the edge represents the relationship in the knowledge graph, which is used to represent a certain connection between different entities.
- **Attributes:** Both entities and relationships in the knowledge graph can have their own attributes.

1.3. Different models based on Transformer

In this section, we briefly describe the three language models we use in our experiments.

1.3.1. BERT

The introduction of Bidirectional Encoder Representations from Transformers (BERT) [2] by the Google AI team in 2018 has marked a milestone in the field of NLP. BERT’s two-way representation, adjusting the context across layers, distinguishes it from other language representation models. BERT’s pre-trained model can be fine-tuned for complex NLP problems such as sentiment analysis and question answering. The model is trained through two phases: pre-training via self-supervised learning on a large corpus, and fine-tuning on task-specific data. Although the architecture remains unchanged, model parameters may differ due to fine-tuning for each task.

1.3.2. RoBERTa

RoBERTa [3] is a modification of the BERT model introduced by Facebook. While the overall structure of RoBERTa is similar to BERT, there are modifications in the pre-training step, including larger batch size, more parameters, and different pre-training techniques. Specifically, RoBERTa removes the Next Sentence Prediction task and implements dynamic

masking instead of static masking. Dynamic masking allows for different masking techniques to be learned through a larger volume of data training for different language representations.

1.3.3. GPT-3

We are using the Text-Davinci-003 model, which is based on the GPT-3 [4] architecture, having been trained on a diverse range of internet text. The model has the ability to generate human-like text, perform natural language processing tasks, and even complete programming code. With its advanced capabilities, Text-Davinci-003 has the potential to revolutionize the field of natural language processing and has numerous potential applications, ranging from customer service and content generation to machine translation and question answering. The model’s large size and diverse training data make it well-suited for a wide range of applications, and its ability to understand and generate text in multiple languages makes it a valuable resource for international businesses and organizations.

1.4. The limitations of current approaches

- The need for extensive corpus for retraining: There have been numerous models based on Transformer architecture, each having a distinct structure, which necessitates a large corpus for pre-training. The pre-training of such vast amounts of data is a computationally demanding task and requires a considerable amount of time.
- Data constraints in fine-tuning step: During the fine-tuning phase for specific domain tasks, if there is a scarcity of datasets, it can result in inadequate model tuning, leading to a reduction in performance.
- Lack of methods to bring in external knowledge: Despite the successes of large language models, such as GPT-3, in performing NLP tasks with limited training data through few-shot and zero-shot learning, there currently lacks a systematic method for incorporating knowledge graphs into these models to enhance their performance.

In this study, we introduce a new method for enhancing text classification by combining knowledge graphs with pre-trained Transformer-based models. We also propose a novel approach using prompts to improve the models’ performance by enabling them to access external knowledge, even with limited training data. Our aim is to improve the classification performance of Transformer-based models when dealing with limited training data. To achieve this, we augment the text representation with a knowledge graph, which provides additional information for model training. We employ three pre-trained Transformer-based models, BERT, RoBERTa, and GPT-3, for text classification tasks. We evaluate the effectiveness of our approach through experiments on two text classification tasks, comparing the results with varying amounts of training data. Our findings indicate that our approach outperforms the conventional method in text classification, even with limited training data.

2. Related Work

Various studies have proposed the combination of Transformer models with knowledge graphs to improve performance on natural language processing tasks. The modifications include a new loss function and bias term based on maximum likelihood estimation [5], the use of semantic information in the knowledge graph to correct relation extraction [6], the incorporation of knowledge graph triples into pre-trained BERT models [7], and the use of a pre-trained Transformer model for knowledge graph construction [8]. Additionally, a question-answering model combined with a knowledge graph was proposed to supplement prior knowledge [9]. These studies demonstrate the effectiveness of combining Transformer models and knowledge graphs for NLP tasks.

Natural language processing encompasses a wide range of tasks, and integrating knowledge graphs with Transformer-based models has been proposed as a solution for various tasks. Lyu [10] combined knowledge graphs and Transformers to address the polysemy problem in Chinese text matching, while Kacupaj [11] proposed a question-answering model that incorporates a graph attention module to explore logical relationships in a text. Yao [12] injected knowledge from knowledge graphs into pre-trained models using the feedforward layer to store knowledge, and Wang [13] dynamically embedded knowledge graphs into language models, embedding both entities and the edges connecting them. Friedman [14] developed a Transformer-based model to extract entities and causal relationships, leveraging BERT and SpERT to generate knowledge graph-like data. These approaches showcase the significance of integrating knowledge graphs with language models for addressing various NLP tasks, with each method providing its unique contribution. Additionally, text matching and question-answering tasks are important NLP tasks, and while pre-trained Transformer models perform well on these tasks, adding knowledge from knowledge graphs can improve their performance. Finally, as the amount of information continues to grow, knowledge graphs can aid in quickly searching for relevant information, making their integration with language models crucial for natural language processing.

In recent years, recommendation systems have become ubiquitous on the internet, facilitating product and search recommendations using natural language processing technology. However, such systems may struggle with understanding user intent or lack sufficient prior knowledge to discern user preferences. To address this issue, Zhou [15] proposes a novel recommendation system combined with a knowledge graph to expand the model’s corpus and improve its recommendation results. Meanwhile, in the context of increasing network intrusions, Ranade [16] develops a GPT2-based model for generating false network intrusion intelligence to improve detection accuracy. Additionally, Zhao [17] presents a method for sentiment analysis by combining prior knowledge from a knowledge graph with a BERT model, which can be used to analyze product positioning. Finally, Do [18] suggests a technique for enhancing question-answering models using a knowledge graph to better classify and answer queries, even when the relevant information is not explicitly mentioned in the text.

3. System Design

In the following section, the data processing, classifier, and technical design of the algorithm are described in detail along with the accompanying pseudocode. The overall workflow is depicted in Figure 1, with the corresponding pseudocode presented below.

Algorithm 1 Text classification

Input: NH, a set of n news headline ($nh_1, nh_2, nh_3, \dots, nh_i, \dots, nh_n$)

Output: Category

```

1: for For each news headline  $nh_i$  in NH do
2:   Split sentences into words as W ( $w_1, w_2, w_3, \dots, w_i, \dots, w_m$ )
3:   for each word  $w_i$  in W do
4:     Do NER task and get the entities from W as E( $e_1, e_2, e_3, \dots, e_i, \dots, e_l$ )
5:     Search on google knowledge graph and get the short description as D( $d_1, d_2, d_3, \dots, d_i, \dots, d_o$ )
6:     Add  $d_i$  in D to  $w_i$ 
7:     Perform Classifier
8:   end for
9: end for

```

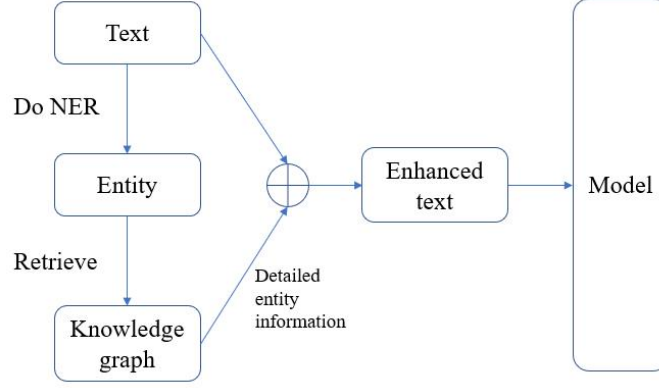


Figure 1. The workflow of our approach

3.1. Approach

A specific example of our data expansion approach is shown in Figure 2.

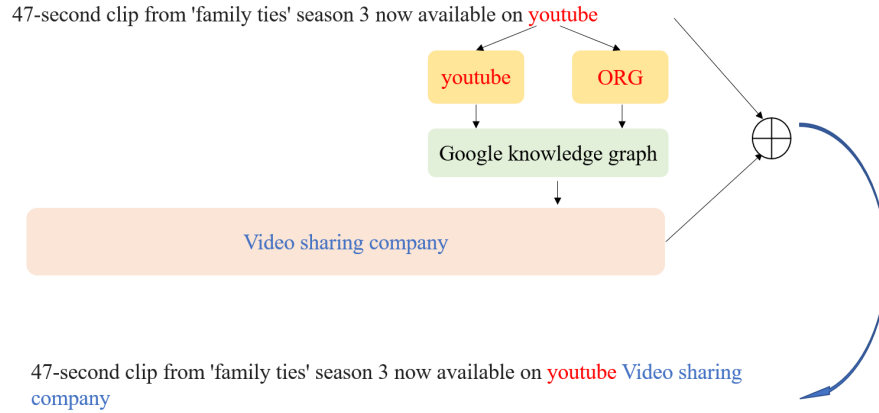


Figure 2. The process of data expansion

3.1.1. Approach for BERT and RoBERTa

To improve the accuracy of text classification in the context of the sarcastic news headline detection dataset, we implement a data expansion technique using an external knowledge graph. The approach involves the following steps:

- Integrating the Google Knowledge Graph as the external knowledge source.
- Identifying entities in the text through named entity recognition technology [19].
- Querying the Google Knowledge Graph for additional information about the recognized entities.
- Augmenting the original text with the description of the corresponding entities retrieved from the Google Knowledge Graph.
- Performing text classification using the BERT and RoBERTa models on the augmented text.

3.1.2. The prompt for GPT-3

We use the GPT-3 model for experiments on the sarcasm detection task. First, we expand the data using the same data expansion method, then for each piece of data to be classified we expand it using the following format: "I'll give you some news headlines and judge whether they're sarcastic. returns 1 if sarcastic, 0 if not sarcastic." Then four randomly selected data in the training set are also added to the prompt, and finally, the data to be classified is placed at the end to make the model understand its task. The prompt method using our data expansion method is the same as that of the original data prompt, the difference is that in the prompt our proposed prompt method adds the description information of the entity. The prompt method for GPT-3 is shown in Figure 3 and Figure 4.

```
I'll give you some news headlines and judge whether they're sarcastic. Returns 1 if sarcastic, 0 if not sarcastic.
Q:a millennial perspective on concur's new app center
A:0
Q:new netflix gas lets users inhale multiple seasons of tv shows
A:1
      ⋮
Q:47-second clip from 'family ties' season 3 now available on youtube
A:
```

Figure 3. The prompt strategy without knowledge graph

```
I'll give you some news headlines and judge whether they're sarcastic. Returns 1 if sarcastic, 0 if not sarcastic.
Q:a millennial perspective on concur Company's new app center
A:0
Q:new netflix Production company gas lets users inhale multiple seasons of tv shows
A:1
      ⋮
Q:47-second clip from 'family ties' season 3 now available on youtube Video sharing company
A:
```

Figure 4. The prompt strategy with knowledge graph

3.2. Datasets

In this section, we briefly describe the two datasets we use in our experiments.

3.2.1. Sarcasm detection task

In our study, we utilize the News Headlines Dataset for Sarcasm Detection [20] as the dataset for our experiment. We selected 1010 of these data for the training set and 425 data for the test set. The data contains the following three attributes.

- is_sarcastic. Shows whether the news is sarcastic. 1 means sarcastic, and 0 means no sarcastic.
- Headline. The headline of the news is provided.
- Link. A link to the original text is given.

3.2.2. News category classification task

In the news category classification task, we used Kaggle's News Category Dataset [21]. This dataset contains news data from 2012 to 2018, which contains 25 categories, namely:

['politics', 'entertainment', 'comedy', 'world news', 'media', 'black voices', 'queer voices', 'sports', 'impact', 'crime', 'weird news', 'business', 'latino voices', 'taste', 'religion', 'travel', 'tech', 'green', 'divorce', 'science', 'arts', 'healthy living', 'education', 'wellness', 'food & drink', 'women']. We selected 2770 of these data as the training set and 924 data as the test set. Each piece of data contains the following six attributes.

- Category. Shows categories of news.
- Headline. The headline of the news is provided.
- Link. A link to the original text is given.
- Short_description. A short summary of the news content.
- Date. The exact date of the press release.

3.3. External knowledge graph

The Google Knowledge Graph [22] is a data representation system that depicts the relationships between various entities. It is a combination of a thesaurus and a lightweight ontology, as each node in the graph is a structured object that includes information. There are tens of thousands of people manually proofreading and editing, and because each person has a unique understanding and mastery of knowledge, the knowledge graph is very informative.

4. Results and Analysis

In order to assess the effectiveness of our proposed method, we conducted two experiments and evaluated the results. The first experiment involved comparing the performance of the original method and our proposed method in the sarcasm detection task. The performance was evaluated using accuracy and F1 score metrics. In the second experiment, we evaluated our proposed method in the news headline classification task, which poses a challenge due to the imbalanced distribution of data among the different categories. The results were analyzed using accuracy and F1 score metrics to provide a comprehensive evaluation of our proposed method.

4.1. Sarcasm detection

In our evaluation of the model’s performance, we focus on the sarcasm detection task. Initially, we use an unmodified model to classify the test data. The results of the training can be seen in Table 1.

Table 1. The performance of two methods(no training)

	<i>Accuracy</i>	<i>F1 score</i>	<i>Precision</i>	<i>Recall</i>
BERT	0.374	0.286	0.683	0.507
RoBERTa	0.362	0.266	0.181	0.500
BERT with knowledge graph	0.372	0.282	0.899	0.899
RoBERTa with knowledge graph	0.362	0.266	0.500	0.266

The performance of the unfine-tuned(zero-shot learning) model on sarcasm detection was observed, and it was found to perform poorly on both the original and the augmented datasets. This is attributed to the lack of fine-tuning, as the pre-training tasks of models like BERT and RoBERTa do not focus on text classification, thus making it challenging for them to perform well on downstream tasks. This experiment highlights the importance of fine-tuning pre-trained models like BERT and RoBERTa in order to achieve optimal performance on specific tasks.

In the following experiment, we investigate the performance of the model using a limited number of training data, specifically 505 samples. The objective of this experiment is to evaluate the effectiveness of our proposed data expansion method in improving the performance of the model when there is a scarcity of task-specific training data.

Table 2. The performance of two methods(505 rows of training data)

	<i>Accuracy</i>	<i>F1 score</i>	<i>Precision</i>	<i>Recall</i>
BERT	0.776	0.726	0.797	0.713
RoBERTa	0.706	0.593	0.751	0.607
BERT with knowledge graph	0.816	0.798	0.804	0.793
RoBERTa with knowledge graph	0.812	0.773	0.838	0.756

The results of using 505 pieces of training data to train the BERT and RoBERTa models for sarcasm detection are displayed in Table 2. The accuracy of the BERT model on the original dataset is 0.776, and the accuracy of the RoBERTa model is 0.706. When the same models were trained on the augmented dataset, their accuracy improved to 0.816 and 0.812, respectively. These results demonstrate that our proposed method of combining Transformer-based models with a knowledge graph effectively improves the performance of the models even with a limited amount of training data.

In the next experiment, we increased the amount of training data to 1010 samples. The results are presented in Table 3.

Table 3. The performance of two methods(1010 rows of training data)

	<i>Accuracy</i>	<i>F1 score</i>	<i>Precision</i>	<i>Recall</i>
BERT	0.819	0.808	0.803	0.814
RoBERTa	0.715	0.626	0.735	0.628
BERT with knowledge graph	0.864	0.847	0.865	0.836
RoBERTa with knowledge graph	0.882	0.872	0.873	0.871

The experimental results indicate that when using the original text, the accuracy of the BERT and RoBERTa models is 0.819 and 0.715, respectively. However, when using the augmented text, the accuracy of the BERT and RoBERTa models increases to 0.864 and 0.882, respectively. The findings from this experiment provide further evidence to support our assertion that incorporating knowledge graphs can lead to better integration of external information and enhance the model’s performance in specific tasks.

4.2. News categories classification

To further investigate the effectiveness of our proposed method in classifying short texts with low data, we conducted experiments on a different classification task - News Categories Classification. This task has a larger number of categories with a significant difference in the number of data samples between them. Some categories have more than a thousand samples while others have only a few. This makes the classification task more challenging for the model, especially for categories with small amounts of data. The experimental results are presented in Table 4, and the ROC curves for the BERT and RoBERTa models with and without knowledge graphs are shown in Figures 5 and 6. The different colored lines in the figures represent different news categories.

Based on the experimental results presented in Table 4, Figures 5 and 6, it is evident that the introduction of a knowledge graph as a data expansion method can significantly improve the classification performance of the models on the News categories classification task.

Table 4. The performance of two methods

	<i>Accuracy</i>	<i>F1 score</i>	<i>Precision</i>	<i>Recall</i>	<i>AUC</i>
BERT	0.679	0.143	0.182	0.150	0.826
RoBERTa	0.738	0.180	0.179	0.192	0.846
BERT with knowledge graph	0.734	0.242	0.258	0.239	0.901
RoBERTa with knowledge graph	0.778	0.286	0.300	0.285	0.911

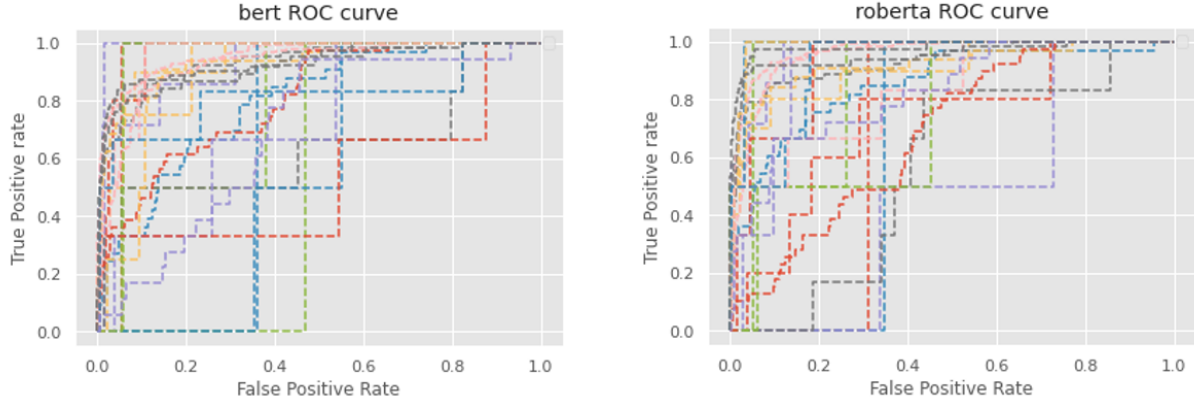


Figure 5. The ROC curve of BERT and RoBERTa without knowledge graph (ROC of BERT on the left, ROC of RoBERTa on the right). The different colored lines in the figures represent different news categories.

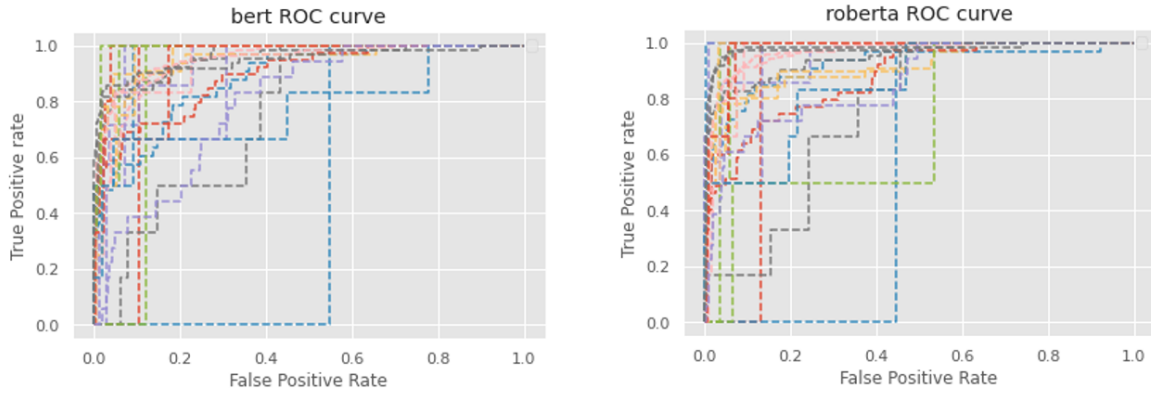


Figure 6. The ROC curve of BERT and RoBERTa with knowledge graph (ROC of BERT on the left, ROC of RoBERTa on the right). The different colored lines in the figures represent different news categories.

Specifically, the BERT and RoBERTa models without the knowledge graph achieved accuracy rates of only 0.679 and 0.738, respectively. However, after incorporating the knowledge graph for data expansion, the accuracy rates of the models improved to 0.734 and 0.778, respectively. This improvement can be attributed to the ability of the knowledge graph to

provide additional external information to the models during training, which enables the models to gain a deeper understanding of the semantic information of different sentences through the attention mechanism. Consequently, the models are better able to incorporate this information in their classification judgments, leading to superior classification performance. In addition, similar to category 3, the model’s performance varies across categories in the classification task. For instance, category 3, comprising 268 samples, achieved AUCs of 0.931 and 0.930 for BERT and RoBERTa without knowledge graphs, respectively. However, introducing knowledge graphs improved the AUCs to 0.952 and 0.945 for BERT and RoBERTa, respectively. This demonstrates that knowledge graph expansion can enhance the model’s discriminative ability by introducing external information, resulting in a more comprehensive fine-tuning of the model. Thus, our proposed approach of combining Transformer models with knowledge graphs can effectively boost the performance of text classification tasks.

4.3. GPT-3 for sarcasm detection

The accuracy of Text-Davinci-003 on sarcasm detection is shown in table 5.

Table 5. The performance of two methods

	<i>Accuracy</i>
GPT-3 without knowledge graph	0.680
GPT-3 with knowledge graph	0.703

By searching the entities in the text in the knowledge graph and adding their descriptions to the text, the model is able to better understand the semantic information of the text and improve its classification ability. The knowledge graph introduces external information to the model, allowing it to learn from a larger and more diverse set of data. By leveraging the knowledge graph, the model is able to expand its understanding of the entities and concepts in the text, and in turn, its ability to classify the text more accurately.

Additionally, the attention mechanism of the Transformer model is another factor that contributes to the effectiveness of this method. The attention mechanism allows the model to focus on the most relevant parts of the text, which is especially important in cases where the text is complex and contains multiple entities and concepts. By integrating the descriptions of the entities into the text, the attention mechanism can better identify the relationships and connections between different parts of the text, ultimately improving the classification performance.

Furthermore, the addition of external information through the knowledge graph also helps the model to overcome the limitations of the training data. In many cases, the amount of labeled data available for training is limited, and the model may not be able to fully capture the nuances and complexities of the language. By leveraging external information from the knowledge graph, the model can effectively supplement the training data and achieve better generalization performance.

5. Conclusions

In this study, we conducted experiments and analysis on two natural language processing tasks, namely sarcasm detection and news category classification, using the BERT, RoBERTa, and GPT-3 models, and achieved promising results. The objective of the sarcasm detection experiment was to compare the performance of the models trained with and without the use of external knowledge graphs on datasets of varying sizes. The goal was to

ascertain whether the introduction of external knowledge graphs could enhance the performance of the models on tasks involving low amounts of data. The results revealed that the model that was trained with the augmented data using knowledge graphs performed better than the model without knowledge graphs when the amount of data was limited.

In the second experiment, we performed the news category classification task, where the dataset included categories with varying amounts of data. To address the challenge of the varying amount of data in each category, we introduced external knowledge graphs to augment the data and evaluated the performance of the models with and without the external knowledge graphs. The results showed that the augmented data with the knowledge graph led to better model performance. This demonstrated that the Transformer model combined with the external knowledge graph could improve the performance of the model in text classification tasks.

The introduction of an external knowledge graph was found to enhance the model’s ability to acquire external information during training, which allowed the model to gain a better understanding of the semantic information of different sentences through the attention mechanism. This improvement enabled the model to acquire more information in classification judgment and achieve better classification results.

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019, pp. 4171–4186.
- [3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. “Roberta: A Robustly Optimized BERT Pretraining Approach”. In: *arXiv preprint arXiv:1907.11692* (2019).
- [4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. “Language models are few-shot learners”. In: (2020), pp. 18702–18712.
- [5] Y. Zhang, X. Shi, S. Mi, W. Zhang, and Q. Huang. “Image captioning with transformer and knowledge graph”. In: *Pattern Recognition Letters* 143 (2021), pp. 43–49.
- [6] W. Huang, Y. Mao, Z. Yang, X. Liu, and W. Wang. “Relation classification via knowledge graph enhanced transformer encoder”. In: *Knowledge-Based Systems* 206 (2020), p. 106321.
- [7] W. Liu, P. Zhou, Z. Zhao, X. Huang, B. Liu, and H. Chen. “K-bert: Enabling language representation with knowledge graph”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 03. 2020, pp. 2901–2908.
- [8] L. Yao, C. Mao, and Y. Luo. “KG-BERT: BERT for knowledge graph completion”. In: *arXiv preprint arXiv:1909.03193* (2019).
- [9] C. Liang, J. Yang, and X. Fu. “Knowledge Graph Enhanced Transformer for Generative Question Answering Tasks”. In: *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I* 30. Springer. 2021, pp. 267–280.
- [10] B. Lyu, L. Chen, S. Zhu, Y. Zhou, and L. Guo. “Let: Linguistic knowledge enhanced graph transformer for chinese short text matching”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 15. 2021, pp. 13498–13506.
- [11] E. Kacupaj, J. Plepi, K. Singh, et al. “Conversational question answering over knowledge graphs with transformer and graph attention networks”. In: *arXiv preprint arXiv:2104.01569* (2021).
- [12] Y. Yao, S. Huang, L. Dong, and X. Liu. “Kformer: Knowledge Injection in Transformer Feed-Forward Layers”. In: *Natural Language Processing and Chinese Computing: 11th CCF*

- International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*. Springer. 2022, pp. 131–143.
- [13] Q. Wang, P. Huang, H. Wang, and L. Wang. “Coke: Contextualized knowledge graph embedding”. In: *arXiv preprint arXiv:1911.02168* (2019).
 - [14] S. E. Friedman, I. H. Magnusson, and S. M. Schmer-Galunder. “Extracting Qualitative Causal Structure with Transformer-Based NLP”. In: *arXiv preprint arXiv:2108.13304* (2021).
 - [15] K. Zhou, W. X. Zhao, J. Bian, X. Li, X. Wang, and D. Yin. “Improving conversational recommender systems via knowledge graph based semantic fusion”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1006–1014.
 - [16] P. Ranade, A. Piplai, S. Mittal, and et al. “Generating fake cyber threat intelligence using transformer-based models”. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–9.
 - [17] A. Zhao and Y. Yu. “Knowledge-enabled BERT for aspect-based sentiment analysis”. In: *Knowledge-Based Systems* 227 (2021), p. 107220.
 - [18] P. Do and T. H. V. Phan. “Developing a BERT based triple classification model using knowledge graph embedding for question answering system”. In: *Applied Intelligence* 52.1 (2022), pp. 636–651.
 - [19] *Jean-Baptiste/roberta-large-ner-english*. URL: <https://huggingface.co/Jean-Baptiste/roberta-large-ner-english/tree/main>.
 - [20] R. Misra and P. Arora. “Sarcasm Detection using Hybrid Neural Network”. In: *arXiv preprint arXiv:1908.07414* (2019).
 - [21] R. Misra. *News Category Dataset*. June 2018. DOI: [10.13140/RG.2.2.20331.18729](https://doi.org/10.13140/RG.2.2.20331.18729).
 - [22] *Google Knowledge Graph Search API*. URL: <https://cloud.google.com/enterprise-knowledge-graph/docs/search-api>.