applied sciences

Article

Empowering Large Language Models to Leverage Domain-Specific Knowledge in E-Learning

Ruei-Shan Lu [1], Ching-Chang Lin [2,*] and Hsiu-Yuan Tsao [3]

[1] Department of Management Information System, Takming University of Science and Technology, Taipei City 114, Taiwan; rslu@takming.edu.tw
[2] Department of Business Administration, Taipei City University of Science and Technology, Taipei City 112, Taiwan
[3] Department of Marketing, National Chung Hsing University, Taichung City 402, Taiwan; jodytsao@dragon.nchu.edu.tw
* Correspondence: cclin@ba.tpcu.edu.tw

Abstract: Large language models (LLMs) have demonstrated remarkable capabilities in various natural language processing tasks. However, their performance in domain-specific contexts, such as E-learning, is hindered by the lack of specific domain knowledge. This paper adopts a novel approach of retrieval augment generation to empower LLMs with domain-specific knowledge in the field of E-learning. The approach leverages external knowledge sources, such as E-learning lectures or research papers, to enhance the LLM's understanding and generation capabilities. Experimental evaluations demonstrate the effectiveness and superiority of our approach compared to existing methods in capturing and generating E-learning-specific information.

Keywords: LLM; domain-specific knowledge; E-learning

check for updates

Citation: Lu, R.-S.; Lin, C.-C.; Tsao, H.-Y. Empowering Large Language Models to Leverage Domain-Specific Knowledge in E-Learning. *Appl. Sci.* **2024**, *14*, 5264. https://doi.org/10.3390/app14125264

Academic Editors: Ahmed Rafea and Julian Szymanski

Received: 2 May 2024
Revised: 9 June 2024
Accepted: 16 June 2024
Published: 18 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the rapidly evolving landscape of artificial intelligence, large language models (LLMs) have emerged as powerful tools, exemplified by ChatGPT's prowess in diverse linguistic tasks ranging from article writing to code generation. Built upon the Generative Pre-training Transformer (GPT) architecture, ChatGPT 4.0 offers a glimpse into the capabilities of LLMs when repurposed for industry applications [1].

While foundational language models demonstrate remarkable proficiency in general tasks, their lack of domain-specific knowledge often limits their utility in specialized domains such as E-learning [2]. Addressing this limitation is crucial for enhancing the performance and applicability of LLMs in industry contexts.

Large language models mark a significant milestone in AI applications, promising transformative impacts across various sectors. In the domain of education, they hold immense potential to revolutionize E-learning systems, offering avenues for personalized learning experiences, intelligent tutoring, and dynamic content generation [3].

Despite the promise, integrating domain-specific knowledge into LLMs remains a formidable challenge. This paper aims to bridge this gap by exploring methods to augment LLMs with domain-specific knowledge for E-learning applications. By doing so, we seek to not only enhance the performance of LLMs but also unlock new possibilities for personalized, effective, and scalable E-learning solutions.

Our research begins with an extensive review of existing literature, focusing on methodologies for training LLMs with domain-specific knowledge and identifying potential benchmarks to evaluate their performance in E-learning contexts. Subsequently, we present two real-world case studies to validate the effectiveness of our approach in generating E-learning content tailored to specific domains.

Through this study, we aim to contribute to the advancement of E-learning technologies by leveraging the capabilities of LLMs enriched with domain-specific knowledge. By highlighting the transformative potential of this approach, we underscore the significance of integrating cutting-edge AI technologies into educational practices.

ChatGPT has successfully captured the public's attention with its wide-ranging language capability. Shortly after its launch, the AI chatbot performed exceptionally well in numerous linguistic tasks, including writing articles, poems, code, and lyrics. Built upon the Generative Pre-training Transformer (GPT) architecture, ChatGPT provides a glimpse of what large language models (LLMs) are capable of, particularly when repurposed for industry use cases [1].

Language models (LMs) are widely used in natural language processing tasks, providing a foundation for many applications. While foundational models can perform remarkably well in a broader context, they lack the domain-specific knowledge to be helpful in most industrial or business applications. LLMs trained on large-scale generic datasets often lack domain-specific knowledge, leading to suboptimal performance in domain-specific applications such as E-learning. To address this challenge, there is a need to enhance LLMs with domain-specific knowledge to improve their performance and utility in industry [2].

Large language models marked an important milestone in AI applications across various industries. LLMs fuel the emergence of a broad range of generative AI solutions, increasing productivity, cost-effectiveness, and interoperability across multiple business units and industries. One of the potential areas is education; LLMs will reform E-learning systems in multiple ways, enabling fair learning and better knowledge accessibility. Educators can use custom models to generate learning materials and conduct real-time assessments [3]. Based on the progress, educators can personalize lessons to address the strengths and weaknesses of each student. In the field of E-learning, utilizing LLMs can bring about transformative changes. Intelligent tutoring systems, automated grading, and E-learning content generation are just a few potential applications where LLMs equipped with domain-specific knowledge can play a vital role. However, integrating domain-specific information into LLMs remains a significant challenge. Even though empowering LLMs in the context of E-learning has attracted widespread attention, there are few relevant benchmarks available to measure the effectiveness of LLMs with specific domain knowledge in the context of E-learning.

In this study, firstly, we will explore existing literature related to the methods of training LLMs with specific domain knowledge and potential benchmarks to measure the performance of the generated content in the context of E-learning. Then, based on potential benchmark variables, we will examine two real cases in the content of E-learning to validate the performance and effectiveness of generated content in the context of E-learning.

2. Literature Review

2.1. Domain-Specific LLM

A domain-specific LLM is a general model trained or fine-tuned to perform well-defined tasks dictated by organizational guidelines [4]. Unlike a general-purpose language model, domain-specific LLMs serve a clearly defined purpose in real-world applications. Such custom models require a deep understanding of their context, including product data, corporate policies, and industry terminologies. Notably, not all organizations find it viable to train domain-specific models from scratch. In most cases, fine-tuning a foundational model is sufficient to perform a specific task with reasonable accuracy. This approach requires fewer datasets, computation, and time.

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and humans through natural language. Retrieval-augmented generation (RAG) is a method that combines the strength of pre-trained models and information retrieval systems [5]. Retrieval-augmented generation is a technique in NLP that allows LLMs like ChatGPT to generate customized outputs that are outside the scope of the data it was trained on. This approach uses embeddings to enable language models to perform context-specific tasks such as question answering. Embeddings are numerical representations of textual data, allowing the latter to be programmatically queried and retrieved [6]. When implemented, the model can extract domain-specific knowledge from data repositories and use them to generate helpful responses. This is useful when deploying custom models for applications that require real-time information or industry-specific context. Retrieval-Augmented Generation is a great technique that combines the strengths of language models like GPT-3 with the power of information retrieval. By enriching the input with context-specific information, RAG enables language models to generate more accurate and contextually relevant responses. In enterprise use cases where fine-tuning might not be practical, RAG offers an efficient and cost-effective solution to provide tailored and informed interactions with users. For example, financial institutions can apply RAG to enable domain-specific models capable of generating reports with real-time market trends [7].

2.2. Benchmark of Generated Content

BERTScore is a widely used metric in NLP for evaluating the semantic similarity between generated text and reference text. It leverages pre-trained BERT models to assess the quality of generated text, including summaries, translations, and responses to questions [8]. Research demonstrates that BERTScore outperforms traditional metrics like BLEU and ROUGE in capturing semantic similarity. It measures the similarity between the generated and reference text in terms of semantic content, yielding Precision, Recall, and F1 Score. Higher values of these scores indicate a stronger degree of semantic overlap, making BERTScore valuable for evaluating the quality of generated text [9]. Additional considerations may include averaging scores across multiple reference texts for a more comprehensive evaluation.

BLEU is a metric for evaluating the quality of text that has been machine-translated from one language to another. It compares the n-grams of the candidate translation with the n-grams of one or more reference translations.

ROUGE is a set of metrics for evaluating automatic summarization and machine translation models by comparing their output against reference summaries or translations.

3. Research Methodology

Our approach involves pre-training the LLM, Llama 2, which was released by Meta AI (New York, NY, USA) in July 2023, and utilizing E-learning materials, including textbooks and research papers, as external sources of knowledge.

Using RAG and source knowledge, we retrieve relevant information from an external data source, augment our prompt with this additional source knowledge, and feed that information into the LLM [10]. The detailed procedures will be presented in the following.

A RAG system consists of three primary parts: the retrieval, augmentation, and the generate.

3.1. Retrieval

Based on the prompt, retrieve relevant knowledge from a knowledge base. In most cases, your “knowledge base” consists of vector embeddings stored in a vector database like Pinecone, and your “retriever” will embed the given input at runtime and search through the vector space containing your data to find the top K most relevant retrieval results, then rank the results based on relevancy (or distance to your vectorized input embedding. The retriever ranks documents based on their locational proximity in a multidimensional vector space, enabling an understanding of textual relationships and relevance between an input and the document corpus.

The retriever is responsible for searching through the knowledge base for the most relevant pieces of information that correlate with the given input, which is referred to as retrieval results.

3.2. Augmentation

Combine the retrieved information with the initial prompt.

This can be done with a prompt formatted to the specific application. For instance, we can declare the following format:

```
"Answer the customers's prompt based on the following context:
==== context: {document title} ====
{document content}

...

prompt: {prompt}"
```

This format can then be used, along with whichever document was deemed useful, to augment the prompt. This augmented prompt can then be passed directly to the LLM to generate the final output.

3.3. Generate

Pass the augmented prompt to a large language model, generating the final output.

On the other hand, the generator utilizes these retrieval results to craft a series of prompts based on a predefined prompt template to produce a coherent and relevant response to the input.

3.4. Experimental Design

To evaluate the effectiveness of our approach, we conduct comprehensive experiments on a large-scale E-learning dataset. We compare our approach with baseline methods, including traditional LLMs and domain adaptation techniques. We measure the model's performance on various E-learning-specific tasks, such as question answering, passage generation, and summarization. Additionally, we assess the model's ability to handle new, unseen E-learning content.

3.5. Baselines

We leveraged LLMs, namely GPT-4, as the backbone to output the response by taking the extra information from either the data retrieval methods or our approach. As for the data retrieval methods, please refer to Figure 1. These methods were employed to retrieve the top three relevant information chunks from Google Cloud which were then used as supplementary information for the backbone LLMs during answer generation.

We make the below baselines.

1. Raw LLM (LLM):

   Questions were directly posed to the backbone LLMs without providing any domain-specific knowledge. In this study, the textbook used is "Digital Marketing."

2. LLM along with DSK (domain-specific knowledge):

   The LLM utilized the domain knowledge from our domain-specific LM as extra information to generate answers.

3. Evaluation Metrics:

   BERTScore is a metric used to evaluate the semantic similarity between two pieces of text, typically a reference sentence and a generated sentence. Therefore, we utilize BERTScore to assess the semantic overlap between the generated answers and the ground truth.

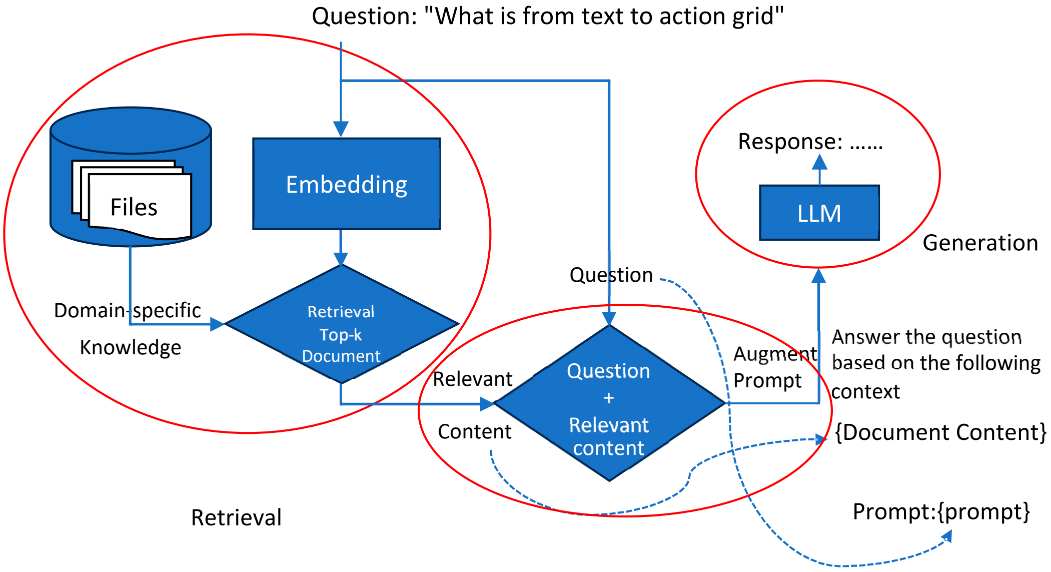
Question: "What is from text to action grid"
Files
Embedding
Domain-specific Knowledge
Retrieval Top-k Document
Retrieval
Question
Relevant Content
Question + Relevant content
Augment Prompt
Response: ……
LLM
Generation
Answer the question based on the following context
{Document Content}
Prompt:{prompt}

Figure 1. LLM + Domain-Specific knowledge + Retrieval Augment Generation.

3.6. Case Study

1. Cases 1 and 2:

We adopt two topics, Innovation Disruption and Sentiment Analysis, extract the content from the textbook "Digital Marketing", and for those questions, ground answers, and generated content from LLM ChatGpt4 and LLM + DSK, please refer to Appendix A. In the following, we attempt to compare the BERTScore calculated from the retrieved chunks from LLM ChatGpt4 and our domain knowledge generated from the domain-specific LM with ground answers.

2. Cases 3 and 4:

We adopt two topics, From Text to Action Grid and Scale-Directed Text Analysis, extract the content from the research paper "From Mining to Meaning" [11], and for those questions, ground answers, and generated content from LLM ChatGpt4 and LLM + DSK, please refer to Appendix A. In the following, we attempt to compare the BERTScore calculated from the retrieved chunks from LLM ChatGpt4 and our domain knowledge generated from the domain-specific LM with ground answers.

4. Results and Analysis

Our experimental results demonstrate the superiority of our approach in incorporating domain-specific knowledge into LLMs for the textbook "Digital Marketing" and research papers "From Mining to Meaning" for E-learning. The enriched LLM outperforms baseline models in terms of accuracy, fluency, and coherence. Moreover, the integration of the knowledge graph significantly enhances the model's understanding and generation capabilities, leading to more informative and contextually relevant responses.

The open-source BERTScore Python package, leveraging Hugging Face's Transformers Library, is used to automate the calculation of BERTScore (Precision, Recall, and F1 Score). The results are shown in Table 1. As for the meaning of Precision, Recall, and F1 Score, please refer to Appendix B.

Table 1. BERTScore of Precision, Recall, and F1 Score.

	Precision	**Recall**	**F1 Score**
Textbook			
Case 1			
LLM ChatGPT4	0.5361	0.5571	0.5464
LLM DSK	0.5954	0.5761	0.5856
Case 2			
LLM ChatGPT4	0.6136	0.5578	0.5844
LLM DSK	0.6269	0.5634	0.5934
Research Paper			
Case 3			
LLM ChatGPT4	0.5539	0.4949	0.5227
LLM DSK	0.7111	0.6312	0.6688
Case 4			
LLM ChatGPT4	0.5243	0.5179	0.5211
LLM DSK	0.6112	0.5745	0.5923

For cases 1 and 2 of testing domain-specific knowledge, ground knowledge and answers are from the textbook of “Digital Marketing”. We employee the topics of Innovation Disruption and Sentiment Analysis. Basically, the semantic overlap metrics of the raw LLM ChatGPT4 and local LLM + DSK are all quite good. However, the BERTScore values of local LLM + DSK are still slightly better than that of raw LLM.

As for cases 3 and 4 of testing domain-specific knowledge, ground knowledge and answers are from a specific research paper “From Mining to Meaning”. Apparently, the score of semantic overlap metrics of the raw LLM ChatGPT4 are far behind the local LLM + DSK.

Finding

The data-retrieved chunks contain scattered information related to keywords in the question. Thus, the local domain knowledge provides a direct answer or response related to those questions if the related keywords can be found in the local domain knowledge. Otherwise, the LLM will seek the pre-train knowledge in the local knowledge.

As for cases 1 and 2, the performance of semantic overlap of raw LLM and LLM with local domain knowledge are all good, the main reason being that the topics Innovation Disruption and Sentiment Analysis are commonly seen concepts and knowledge over the Internet. Thus, the score of semantic overlap is all very good, regardless of whether it was generated from raw LLM or LLM plus local domain knowledge.

On the contrary, in cases 3 and 4, the performance of semantic overlap of LLM with local domain knowledge is outperformed compared to raw LLM. The main reason is that the topics “from text to action grid” and “scale-directed text analysis” are all newly published papers and quite specific knowledge instead of commonly known concepts and knowledge. Thus, obviously, the score of semantic overlap of LLM with local domain knowledge performs better than raw LLM.

5. Conclusions

This paper presents a novel approach to empower large language models (LLMs) with domain-specific knowledge in the E-learning domain. Our method combines pretraining LLMs with the integration of external, domain-specific knowledge sources to enhance their performance in E-learning applications. Experimental results demonstrate the effectiveness and superiority of our approach in capturing and generating E-learning-specific information, especially for knowledge and concepts that are not commonly used or widely known, such as newly published research papers.

The practical implications of our approach are significant. By integrating domain-specific knowledge into LLMs, our method can improve the accuracy and relevance of intelligent tutoring systems, automated grading, and E-learning content generation. This can lead to more personalized and effective learning experiences, enhancing educational outcomes for students.

We also conducted a thorough analysis of the strengths and limitations of our approach. Among its strengths, our method offers improved specificity and relevance in E-learning contexts, making it particularly useful for handling up-to-date information and niche topics within the domain. Additionally, the integration of external knowledge sources ensures that the LLMs remain current with the latest developments in the field.

However, our approach does have limitations. One major limitation is the dependency on the continuous updating of external knowledge sources to maintain the LLMs' relevance. This can be resource-intensive and may require significant effort to ensure that the knowledge base remains comprehensive and current. Furthermore, the process of integrating external knowledge sources can be complex and may require specialized expertise.

To address these limitations and enhance the applicability of our approach, we propose several avenues for future research. Firstly, exploring automated methods for continuously updating and expanding the domain-specific knowledge base will be crucial. Secondly, further research into in-context learning and the design of carefully formatted prompts can improve the ability of LLMs to retrieve and utilize relevant documents accurately. This includes developing strategies for LLMs to efficiently learn and adapt to new information without extensive retraining.

In conclusion, our approach shows considerable promise in advancing the capabilities of LLMs in E-learning scenarios. By addressing its limitations through targeted research and development, we can maximize its potential and ensure its effectiveness across a broader range of applications within the educational domain.

Future research to enhance large language models (LLMs) in the E-learning domain should focus on automated systems for continuously updating domain-specific knowledge bases, adaptive prompt design for improved context relevance, and efficient learning techniques like few-shot learning to minimize retraining. Additionally, integrating diverse multimodal data sources, developing scalable infrastructure, and creating personalized learning pathways can further enhance their utility. Addressing ethical concerns, such as bias mitigation and data privacy, alongside exploring the ethical use of AI in education, will ensure the responsible deployment of these advanced models. These efforts can maximize the effectiveness of LLMs, leading to more accurate, relevant, and personalized educational experiences.

Author Contributions: Conceptualization, R.-S.L. and H.-Y.T.; Methodology, R.-S.L. and H.-Y.T.; Software, R.-S.L. and C.-C.L.; Formal analysis, C.-C.L.; Writing—original draft, H.-Y.T.; Writing—review & editing, C.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Case 1 Textbook Disruptive Innovation

- Question:

  Please explain disruptive innovation and give an example to illustrate it.

- Ground Answer:

  Disruptive innovation.

Then there was a lot of competition from other industries in the current industry, such as the emergence of MP3 online music sharing software (Napster), which caused the music industry to disintegrate. Netflix's shift from online DVD lending to streaming and homemade films has led to two disruptive innovations, the first being competition from different operating models in the same industry, and the second being a challenge to the existing film and television industry, including production, distribution and cinemas.

Due to factors such as the evolution of technology and the introduction of disruptive innovation business models, the abundant capital accumulated by enterprises over a long period of time is very likely to collapse in an instant in three to five years due to a lack of innovation, or failure to directly look at and solve the problems encountered inside and outside the enterprise. Nokia and Kodak are well-known examples of the kind of people who once dominated the industry, but eventually faced bankruptcy or annexation.

Therefore, in this era of dynamic competition, if you want to maintain your advantage for a long time, you must obtain a short-term advantage in every short-term competition through continuous small innovation, flexible organizational structure, and dynamic strategic planning, and at the right time, you will give up the previous advantages and prepare for the arrival of the next short-term advantage. Before the previous advantage fades, look for another advantage that will continue to survive and maintain a dynamic competitive advantage.

Appendix A.1.1. Generated Content from LLM ChatGPT4

Disruptive innovation is a concept coined in 1997 by Harvard Business School professor Clayton Christensen that describes a pattern of innovation that typically emerges in a market and erodes the traditional market leader by offering a relatively cheap, simple, easy-to-use product or service. Disruptive innovation not only changes the market landscape but also opens up new market opportunities that are often overlooked or undervalued at first.

Here is an example of a disruptive innovation:

Case in point: Disruptive innovation in online streaming services.

In the past, people habitually bought or rented DVDs, VHS tapes, or Blu-ray discs to watch movies and TV shows. It is a market dominated by big movie rental stores like Blockbuster, who have a large number of brick-and-mortar storefronts for people to rent video entertainment.

However, with the popularity of the Internet and the increase in internet speeds, online streaming services (such as Netflix, Hulu, Amazon Prime Video) have emerged, and these services have provided a whole new way to watch movies. These online streaming platforms allow users to watch a large number of movies and TV shows on a variety of devices anytime, anywhere, with a subscription or rental, without having to buy or rent physical media. What is more, they are often offered at a relatively low price and offer personalized recommendations to improve the user experience.

This model of online streaming services is characterized by disruptive innovations:

1. Low cost: The price of these services is usually cheaper than buying a DVD or renting it, so it attracts a wide range of users.
2. Convenience: Viewers can access entertainment content anytime, anywhere, without having to travel to a physical store or wait for mail service.

3. New markets: Online streaming services have opened up a whole new market and attracted a large number of users who would otherwise not be attracted by traditional rental stores.

Over time, online streaming services have not only disrupted traditional video rental stores, but have also built a huge user base across the globe and become a huge entertainment industry. This example highlights how disruptive innovation can revolutionize the market landscape and create entirely new business opportunities, while also warning us about the possibility that current market leaders may be replaced by competitors from emerging technologies or business models.

Appendix A.1.2. Generated Content from LLM + DSK

Please explain disruptive innovation and give an example to illustrate it.

Disruptive innovation is a concept coined by Harvard Business School professor Clayton M. Christensen to refer to the emergence of new technologies or business models and how they can gradually replace existing products or services in the market, thereby disrupting traditional industries.

Disruptive innovation often fails to meet the needs of traditional markets in the early stages, but as technology continues to advance and costs decrease, new products or business models begin to meet the needs of more and more consumers, and thus defeat the previously dominant product or service. This process is often carried out in a low-cost, simplified manner, and disruptive innovation establishes new rules of competition in the marketplace.

A classic example of disruptive innovation is the impact on the traditional music industry. When music in MP3 format began to mature and the Internet became popular, Napster, an online music sharing service, emerged in 1999, allowing users to download music for free over the Internet. This has had a huge impact on the traditional record industry, as they cannot generate revenue in the traditional way.

Another example of disruptive innovation is Netflix's challenge to the film and television industry. At first, Netflix was an online DVD rental platform, but with the popularity of the Internet and the advancement of video streaming technology, Netflix began to offer video streaming services and make its own titles. This disruptive innovation has changed the way people watch videos, eliminating the need for physical DVDs and the need to rely on traditional movie theaters and television stations.

These disruptive innovations have had a huge impact on existing industries, and many traditional enterprises are at risk of being eliminated. However, at the same time, because of these innovations, new markets and business opportunities have emerged, bringing new competitiveness and opportunities to enterprises. Disruptive innovation has changed the way we live and work, and it has also challenged traditional business models, providing inspiration and reference for future innovations.

Appendix A.2. Case 2 Textbook Sentiment Analysis

- Question:

Please describe the sentiment analysis and sentiment dictionary and its application examples.

- Ground Answer:

Sentiment Analysis, also known as Opinion Mining, refers to the subjective and sentiment assessment of the extraction source through the application of natural language processing technology and text analysis.

Sentiment analysis-related thesaurus.

Sentiment thesaurus is the most critical resource for sentiment analysis, almost all participation needs to rely on a thesaurus of many different emotion words for analysis, common sentiment thesaurus includes AFINN and Bing Liu opinion vocabulary, which also means that sentiment thesaurus plays an indispensable role in sentiment analysis. The following will be explained in detail for the abovementioned thesaurus.

1. AFINN Dictionary of Emotions

It is a rating ranging from −5 (Very negative) to 5 (Very positive) of a list of English words by Finn Årup Nielsen completed by manual marking between 2009 and 2011 (Nielsen, 2011), attached to the polarity (frontal/negative). This thesaurus is ANEW (Affective Norms for English Words) lexicon-based, but AFINN is more focused on microgrids. The language used on online platforms such as Zhi is suitable for analyzing the sentiment of online comments because its word list contains slang, obscene words, acronyms, and Internet buzzwords, which are commonly used after the development of the Internet. AFINN is currently available in two versions, the first being AFINN-96 with a total of 1468 unique words and phrases, and the second edition of AFINN-111 is the latest version with a total of 2477 words and phrases.

2. Bing Liu Thesaurus

Proposed by Hu and Liu in 2004, it mainly distinguishes between positive and negative sentiments of single words, and the positive list includes agile (Agile), Enjoy (Enjoy), Lift (Improving). The negative list has anger (Anger), Reject (Refusing), Incapacitated (Unable). There are 2006 positive and 4783 negative words, tailored specifically for social media, as it takes into account changes in specific areas such as misspellings and slang.

Application of sentiment analysis.

Sentiment analysis can be applied to comments generated on blogs, micro-blogs, forums, chats, and social media to track sentiment related to political elections, finance, and consumer issues, which also reflects the validity and diversity of sentiment analysis research. For example, in the 2014 dispute between Ke P. and Lian Shengwen, through big data, the high correlation between "youth" and "hot dance" was excavated from popular and trending keywords on the Internet. Therefore, arrange a hot dance event, and then use the physical event to find out these groups from the Internet and combine them with the election, and become "Go! Youth! Be Yourself" series of activities. In this way, it is to find out the young people who are apathetic to politics and do not know about Ke P. Use popular keywords from web searches as the subject for developing marketing campaigns and postings.

Recent studies have shown that people express their emotions through online platforms, blogs, social media, etc., which also shows that sentiment analysis has a certain benefit for the development of enterprises. At present, there is an online tool that can analyze Chinese emotions, as mentioned earlier, called Lexalytics. It is very easy to operate, as shown in Figures 2–5, its online URL is https://www.lexalytics.com/demo, and interested people can try it out.

Appendix A.2.1. Generated Content from LLM ChatGPT4

Sentiment analysis is a natural language processing technique designed to identify and evaluate sentiment or sentiment polarity in text. It is generally divided into three main categories: positive, negative, and neutral. Sentiment analysis can be achieved through different methods, one of which is to use a sentiment dictionary.

An emotional dictionary is a vocabulary that contains a word or phrase and the emotional polarity (positive, negative, or neutral) associated with it. These words are labeled as words with emotion based on how they are used in the text. There are two main aspects to the application of emotional dictionaries:

1. Sentiment Analysis: Sentiment dictionaries can be used for sentiment analysis to help machines understand the sentiment content in text. By matching words in the text with words in the sentiment dictionary, a sentiment score for the text can be calculated to determine the sentiment polarity of the text. For example, if a word is marked as "positive" in the sentiment dictionary, the presence of that word in the text may indicate that the text has positive sentiment.

2. Affective Vocabulary Expansion: Affective dictionaries can also be used to expand affective vocabulary. When a sentiment analysis model uses a sentiment dictionary, it can identify known sentiment words but may not be able to process new, unknown sentiment words. In this case, the vocabulary can be expanded using an emotion dictionary to label new words as having sentimental polarity to improve the performance of the model.

Example:

Suppose we have an emotional dictionary with words and emotional polarities. Here is a sample part of a sentiment dictionary:

- "Pleasant": Positive
- "Sad": negative
- "Surprised": Positive
- "Disappointment": Negative
- "Calm": neutral

Now, let us look at an example of a practical application of sentiment analysis:

Text: "The weather today is amazing!"

In this example, our sentiment analysis system will use the sentiment dictionary to match words in the text and analyze sentiment based on the sentiment polarity entries in the sentiment dictionary. For this sentence, it will notice the words "fantastic" and "pleasant", both of which are associated with positive emotions. As a result, the text will be classified as positive sentiment.

The application of sentiment dictionary also includes social media sentiment analysis, sentiment evaluation of product reviews, sentiment classification of news articles, and so on. By using sentiment dictionaries in conjunction with machine learning models, we are able to more accurately analyze the emotional content in the text, allowing us to better understand how people express their emotions in different contexts.

Appendix A.2.2. Generated Content from LLM + DSK

Sentiment analysis is a method of analyzing emotions or emotional tendencies in text through techniques such as natural language processing and machine learning. It can identify and quantify positive, neutral, or negative emotions in text to further understand and analyze the emotional content of the text.

Sentiment dictionaries are a common resource for sentiment analysis, as they collect a large number of words and phrases and label each word with its corresponding emotional polarity (positive, neutral, or negative). There are two main aspects to the application of sentiment dictionaries: the extraction of sentiment words and sentiment analysis.

First of all, the emotion dictionary can be used for the extraction of emotion words. Through the Mood Dictionary, you can quickly and accurately extract emotionally inclined words from the text, which have characteristics that describe emotions or emotions, such as "like", "worried", "sad", etc. This method can be used to analyze user reviews, news articles, and more on social media to understand people's emotional reactions to specific issues.

Second, sentiment dictionaries can be used for sentiment analysis. By calculating the distribution of emotional words in the text and the emotional polarity of the emotional vocabulary, the emotional content of the whole text can be analyzed and quantified. For example, when analyzing users' evaluations of a brand on social media, you can use an emotional dictionary to determine whether the emotional tendencies of the review are positive, neutral, or negative, so as to assess users' attitudes and emotions towards the brand.

Example: Suppose you have an emotional dictionary that contains the positive emotional words "like", "happy", and the negative emotional words "hate", "sad", etc. In response to the following user review "I like this product very much, it makes me happy and satisfied", you can quickly tell that the review has a positive emotional tendency through the emotional dictionary.

Sentiment analysis and sentiment dictionaries have a wide range of applications, and they can be applied to social media monitoring, market research, sentiment analysis, consumer opinion analysis, and other fields to provide important information and insights in business and society.

Appendix A.3. Case 3 The Research Paper (More Specific Domain Knowledge)

- Question:

  What is the from-text-to-action grid.

- Ground Answer:

  From text to strategy: Developing the FTTA grid.

While the metrics gleaned from text analysis can be interesting barometers and trackers of whatever is measured, the metrics alone often do not provide the depth needed to inform firm strategy. This has meant that despite the ease at which textual data can be collected and analyzed, insights from such efforts are often relatively thin. In this section we describe how data on what topics are mentioned in a text (aspect), coupled with data on how frequently those topics are mentioned (attention) and the sentiment they receive (opinion), can provide strategic insight.

We next develop a new framework, the From-Text-To-Action (FTTA) grid, to help marketers better interpret the data that can be gleaned from textual data. The FTTA grid is based on importance–performance analysis, a widely accepted method that is well known for its simplicity and ease of application (Martilla and James, 1977). Importance–performance analysis is based on customer satisfaction being a function of both expectations regarding a specific attribute and judgment of the attribute's performance. Key is the gap between a customer's expectations and their judgment of the performance received during service. Such analysis enables identification of attributes that have the most impact on customer satisfaction and repetitive purchase behavior.

Such information helps marketers both evaluate competitive position and enables prioritization of available strategies to enhance customer satisfaction.

The FTTA grid, based on importance–performance analysis, provides a generalized method for using text-derived metrics to determine which attributes are important to customer service. Given a selected attribute or aspect, the following two components are classified: (1) the importance of the attribute and (2) judgments of its performance. Proxies for these components can be calculated from textual data analytics (see Section 2.3). The significance of an attribute can be calculated based on the level of interest (attention) a specific aspect receives. Similarly, the perceived performance of an attribute can be estimated using the expressed opinion (or sentiment) for that aspect. The standardized data for each aspect's interest and opinion can then be paired and used as coordinates to plot the aspect on a two-dimensional FTTA grid.

Appendix A.3.1. Generated Content from LLM ChatGPT4

As of my last update in January 2022, I am not aware of a specific concept or tool called the "from text to action grid". However, the phrase "from text to action" suggests a progression from understanding or interpreting a written piece of material to taking actions based on that understanding. In education, for instance, this progression could be seen in strategies that move students from reading and understanding a text to applying or acting upon that understanding in some way.

If you are thinking of a tool or concept that has been developed or popularized after 2022, I may not have information on it. Alternatively, it could be a concept or framework specific to a certain field, institution, or context that was not widely recognized outside of that area by 2022.

To help you better, could you provide more context or clarify what you are referring to? If it is a newer or more niche concept, you might need to refer to more recent sources or specific institutions or practitioners associated with the idea.

Appendix A.3.2. Generated Content from LLM + DSK

The From-Text-To-Action (FTTA) grid is a framework that helps marketers analyze textual data and transform them into actionable insights. It is based on the concept of importance–performance analysis, which measures customer satisfaction based on their expectations and judgments of various attributes.

The FTTA grid consists of two axes:

1. Aspect: This axis represents the different topics or attributes that are mentioned in the text. It helps identify what customers or stakeholders are talking about or what is important to them.
2. Attention: This axis represents the frequency or attention that each aspect receives in the text. It helps determine the level of importance or interest in each aspect.

The FTTA grid also takes into account the sentiment or opinion associated with each aspect. By analyzing the aspects, attention, and sentiment, marketers can gain insights into customer perceptions, needs, and preferences. This information can then be used to develop marketing strategies and actions that align with customer expectations and preferences.

Overall, the FTTA grid provides a structured framework for understanding and leveraging textual data to drive marketing actions and decision-making.

Appendix A.4. Case 4 The Research Paper (More Specific Domain Knowledge)

- Question:

  What is scale-directed text analysis

- Ground Answer:

The first step involved in scale-directed text analysis is to generate a dictionary for each dimension of a chosen marketing construct. Words from each dimension's scale items are used as stems from which to automatically build dictionaries using semantically similar thesaurus words. Second, a set of initial data for machine-learning calibration is collected. These data must include paired free-form written text and Likert-scale response data, both of which assess the focal marketing construct. Third, these initial data are randomly divided into training and holdout samples. Text data from the training data set are then analyzed against EJM 54.3 514, the dictionary developed for each of the dimensions of the marketing construct being analyzed. Phrases that match each dimension are compiled and, in a fourth step, are evaluated for sentiment. This information is then used to calculate overall raw scores for each of the evaluated dimensions. Next, these raw text scores are compared against the paired Likert-rating data to develop standard conversion values that equate both scores. Finally, the entire approach is tested using the holdout sample. This involves running the holdout text data through the dictionary and then converting the raw text scores using the conversion values derived from the training data. The resulting EMSS can then be tested against each participant's AMSS (Likert scores). If there is no significant difference between the estimated and AMSS, then the scale-directed text analysis process is considered successful; the technique can then be used on text data alone going forward

Appendix A.4.1. Generated Content from LLM ChatGPT4

As of my last update in January 2022, "scale-directed text analysis" is not a widely recognized or standard term in the fields of text analysis, linguistics, or data science. If it has emerged as a specific methodology or term after this date, I would not have that information.

When we break down the term:

- “Scale” can refer to the size, level, or degree of something.
- “Directed” might suggest a targeted or specific approach.
- “Text analysis” is the process of extracting meaningful information from textual data.

From the components, “scale-directed text analysis” could suggest a type of text analysis that is tailored or adjusted based on the scale (e.g., macro-level trends vs. micro-level details) at which you are examining the data. This is, however, a speculative interpretation without a specific context or source for the term.

If this term is from a specific academic paper, industry report, or specialized context, it would be best to refer directly to that source for a clear definition or explanation. Otherwise, you might also consider reaching out to professionals or researchers in the field of text analysis or natural language processing for clarification.

Appendix A.4.2. Generated Content from LLM + DSK

Scale-directed text analysis is a methodology used to analyze textual data by combining traditional textual and sentiment analysis with the structure and rigor of traditional marketing scales. It aims to gain insight into multiple dimensions of a construct, such as service quality, by analyzing and classifying text against specific construct dimensions.

The process involves the use of machine-learning techniques to automate the development of construct-specific dictionaries based on a small set of seed words. Sentiment analysis techniques are then used to analyze the sentiment expressed in the text and classify it against the construct dimensions.

By using this approach, marketers and researchers can gain a more detailed understanding of the various aspects influencing consumer experiences and evaluations. It goes beyond simple sentiment analysis, which only indicates the overall sentiment, to identify the specific dimensions that contribute to the overall assessment.

Scale-directed text analysis enables the extraction of valuable insights from large volumes of textual data, allowing for more informed decision-making and understanding of consumer preferences and experiences.

Appendix B

BERTScore operates on a range between 0 and 1, with some nuances:

- 0: This signifies no semantic similarity between the candidate and reference texts. In simpler terms, the sentences have no meaning in common.
- 1: This represents a perfect semantic match. The candidate text captures all the meaning of the reference text, potentially even using different wording.

However, interpreting the exact value depends on your specific task. A BERT-F1 score of 0.9 might be excellent for summarizing factual content, while a score of 0.8 might be great for creative writing where capturing the overall feeling is more important than exact details.

Here is a breakdown of the three components that make up the overall BERTScore:

- Precision (BERT-Precision): This measures how well the candidate text avoids introducing irrelevant information. It asks: “Out of the information in the candidate text, how much is actually relevant to the reference text?”
- Recall (BERT-Recall): This measures how well the candidate text captures all the relevant information from the reference text. It asks: “Out of the information in the reference text, how much is included in the candidate text?”
- F1 (BERT-F1): This is a harmonic mean that combines both Precision and Recall, providing a single score between 0 and 1. It offers a balance between the two metrics.

In essence, a score near 1 in BERTScore indicates a high degree of semantic similarity between the candidate and reference texts. The specific meaning depends on the context of your task and what level of detail or creativity is desired.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Thirty-first Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
2. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
3. VanLehn, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **2011**, *46*, 197–221. [CrossRef]
4. Howard, J.; Ruder, S. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018.
5. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020.
6. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 27th Conference on Neural Information Processing Systems (NeurIPS 2013), Lake Tahoe, NV, USA, 5–10 December 2013.
7. Ooi, K.B.; Tan, G.W.H.; Al-Emran, M.; Al-Sharafi, M.A.; Capatina, A.; Chakraborty, A.; Dwivedi, Y.K.; Huang, T.-L.; Kar, A.K.; Lee, V.-H. The potential of Generative Artificial Intelligence across disciplines: Perspectives and future directions. *J. Comput. Inf. Syst.* **2023**, 1–32. [CrossRef]
8. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 26–30 April 2020.
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, MN, USA, 2–7 June 2019.
10. Lin, W.; Byrne, B. Retrieval Augmented Visual Question Answering with Outside Knowledge. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), Abu Dhabi, United Arab Emirates, 7–11 December 2022.
11. Tsao, H.-Y.; Campbell, C.; Sands, S.; Mavrommatis, A. From mining to meaning: How B2B marketers can leverage text to inform strategy. *Ind. Mark. Manag.* **2022**, *106*, 90–98. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.