Data Analysis of Social Media Impact on Brand Perception: RoBERTa Question Answering and RAG

1. Literature Review

Introduction

This literature review explores the intricate dynamics of how brands are perceived in the context of interactions on social media. This study focuses on the use of NLP techniques to extract semantic insights from social media text, the use of pre-trained models for answering questions, and the application of time series analysis to identify patterns in social media engagement over time. This review aims to provide a comprehensive understanding of the complex relationship between social media discourse and brand perception

1. Analyzing the Impact of Social Media on Brand Image

Various studies have emphasized the substantial influence of social media on how brands are perceived. The studies conducted by (Rizwan Ali Khadim, B. Zafar, and M. Younis, 2014) and (L. V. Thang et al., 2016) both concluded that social media communication and engagement have a positive impact on consumer brand perceptions and purchase intentions. This is especially accurate for brands that are focused on technology. The studies conducted by (Helal and Ozuem, 2021) and (Dwairi, Harb and Shehabat, 2020) highlight the significance of social media in promoting brand-customer connections and raising brand recognition. They emphasize the crucial role of electronic word-of-mouth (e-WOM) and user-generated content in facilitating this process. (Ismail, Nguyen and Melewar, 2018) and (Ferreira and Zambaldi, 2019) examine the effects of perceived social media marketing activities on brand and value consciousness. (Ismail, Nguyen and Melewar, 2018) also highlighting the potential for social media to influence materialism and conspicuous consumption. Lastly, (Guida and Wilson, 2017) underscores the role of social media in shaping brand perceptions in the fashion industry, particularly through the development of brand-customer relationships.

Social media are online environments that are created and facilitated by computer-mediated technology. As a result, businesses can utilise these kinds of media as a powerful instrument for marketing and advertising to reach various client segments both locally and internationally and build brand awareness. There is a shortage of data regarding the impact of social media platforms on customers' acquisition of brand awareness. Furthermore, not every business achieves success in this field. Therefore, it is imperative to conduct additional research in this field. This paper offers an opportunity to identify and examine the key factors that may contribute to the emergence of this behavior and proposes a model for understanding such phenomena. The study's findings showed that consumer brand awareness is positively influenced by e-WOM, user-generated content, and product quality. The findings contribute to the comprehension of the capacity of social media campaigns to enhance value and illustrate how the perception of brands is influenced through this novel communication channel.

This study partially addresses the drivers of brand corporate reputation attributable to branding in social media, in particular, the building of community engagement and two of its antecedents. They tested a theoretical framework with brand involvement and perceived homophily as the antecedents of social media community engagement and the relation of the latter with the corporate reputation (Van Doorn et al., 2010; Wirtz et al., 2013). Research suggests that creating a brand community or page on social media improves brand corporate reputation (Dijkmans, Kerkhof and Beukeboom, 2015) increases sales and returns on investments, and fosters positive word-of-mouth (Kumar et al., 2013).

This research investigates the impact of perceived social media marketing activities on brand- and value consciousness. It further examines the effect of social media usage on materialism, brand consciousness and conspicuous consumption, as examining materialism-centric behaviour is becoming important in a consumption-based economy. A self-administered questionnaire was developed and administered to a sample of 346 undergraduate students. Two different research models are tested and confirmed. The findings of this research indicated that perceived social media marketing activities have a significant effect on brand loyalty; brand consciousness and value consciousness mediate the relationship between perceived social media marketing activities and brand loyalty. Moreover, evidence supports the idea that the greater the use of social media, the greater the tendency towards materialism and conspicuous consumption. This study confirms the growing importance of perceived social media marketing activities in envisioning brand loyalty and provides insights into the impact of social media on materialism and conspicuous consumption (Ismail, Nguyen and Melewar, 2018).

The study of (Khan et al., 2023) says whether it is based on fact or fiction, a brand’s image is a crucial part of its overall marketing strategy. A brand image is an association formed in a consumer’s mind when they think of a specific brand. Overall, a brand’s image can be defined as a consumer’s impression or memory of a particular product or service. Social media influencers often send out more product messages to consumers than companies do. By using social media influencers to promote a product, consumers’ perceptions of it changed (Arora and Sanni, 2019). These results align with a study that discovered a positive correlation between the strength of influencers' brand images and the perception of the product. (Chakraborty and Bhat, 2018) argue that social networking influencers can assist brands in developing favorable brand perceptions due to their superior ability to influence consumer behavior and purchase intentions.

However, as time passed, social networking sites evolved into a complex amalgamation of endless opportunities from the fields of artificial intelligence (Chin, Marcelin and Newsted, 2003), cognitive science, machine learning (Djafarova and Rushworth, 2017), deep learning, image processing (Dodoo, 2018), cryptography and network security (Freberg et al., 2011). Customers live in a digital world where almost everything is accessible via a single click or touch. From monstrously large desktop computers to small laptops, palmtops, and now smartphones, humans have advanced toward a century of endless opportunities. One such example is social networking markets, also known as electronic stores or e-commerce, which have drastically changed the way consumers shop now. They have not only transformed the product sales process as a whole but have also changed consumer purchasing habits.

1. Natural Language Processing and Understanding

Natural Language Processing is a branch of machine learning that deals with text and speech. It is a way for computers to analyse, understand, and derive meaning from human language in a smart and useful way. Through the use of NLP, developers can employ techniques to arrange and systematize information in order to accomplish tasks such as automated summarization, translation, identification of named entities, extraction of relationships, analysis of sentiment, recognition of speech, and segmentation of topics. NLP has garnered significant interest for its ability to computationally represent and analyze human language. NLP has found widespread use in diverse domains including machine translation, email spam detection, information extraction, summarization, medical applications, and question answering. To comprehend this computational portrayal of human language, it is crucial to differentiate between tokens and types. A token refers to a specific sequence of characters that is grouped together as a meaningful unit for processing. On the other hand, a type encompasses all tokens that contain the same sequence of characters (Agarwal, 2019).

The study of (Liu et al., 2020), aimed to understand the mental health disparities faced by the transgender community, researchers analysed social media posts to categorize sentiment and build machine learning models. This analysis of social media data has the potential to improve the understanding of the transgender community's well-being and inform interventions to support them.

The methods and difficulties of sentiment analysis are described by (Khan et al., 2016). The biggest obstacle in this field is the dearth of dependable and effective software and resources. Additionally, they suggest enhancing language understanding to enhance knowledge extraction. It covers a wide range of business and social science disciplines. Analysing sentiment is a more recent marketing technique. They employed machine learning to find a solution. They contest that NLP is qualified to respond to any of its queries. Negation is predicated on NLU issues, such as word sense disambiguation, co-reference resolution, and domain awareness. Since sentiment analysis solely examines sentiment, it is a limited NLP topic. With complex network analysis, you can arrange text at random.

Semantic analysis is an essential feature of the NLP approach. It indicates, in the appropriate format, the context of a sentence or paragraph. Semantics is about language significance study. The vocabulary used conveys the importance of the subject because of the interrelationship between linguistic classes.

1. Sentiment Analysis

A lexicon and rule-based sentiment analysis tool, especially tuned to the sentiments expressed on social media, is called VADER (Valence Aware Dictionary and Sentiment Reasoner). VADER, created by (Hutto and Gilbert, 2014), is a demonstration of handling text from social media platforms, including slang, emoticons, hashtags, and abbreviations. Studies show that because VADER is sensitive to both the polarity (positive/negative) and intensity (strength) of emotions, it performs remarkably well on social media datasets.

When applied to Twitter data, for example, the study of (Ranco et al., 2015) has demonstrated that VADER performs better than traditional text analysis algorithms, demonstrating its resilience in capturing the conversational and inventive language characteristic of social media. VADER has been used in several studies to evaluate Twitter data in order to determine public opinion on a variety of subjects, including stock market fluctuations, political developments, and product reviews.

In a notable study, (Ranco et al., 2015) effectively applied VADER to predict trends in the stock market based on sentiment analysis from Twitter, showcasing its usefulness in predictive analytics. VADER is particularly effective for sentiment analysis in these situations because of its domain-specific orientation towards social media text. The extensive lexicon, comprising a vast array of emoticons and internet-specific expressions, enhances its applicability.  
  
  
  
VADER is renowned for its efficacy in analyzing Twitter data, and it is also extensively utilized in other social media platforms such as Facebook, Instagram, and online forums. This review examines comparative analyses that evaluate the efficacy of VADER on different social media platforms. Research comparing sentiment analysis tools across different platforms suggests that VADER consistently demonstrates strong performance. This is partly due to its ability to comprehend negations and modifiers that are context-specific, which are frequently encountered in informal online communication (Severyn and Moschitti, 2015). The robustness of VADER across various social media platforms confirms its utility as a versatile tool in sentiment analysis, capable of handling diverse data sources without significant loss in accuracy.

On the other hand VADER has limitations in dynamic social media environments where new slang and expressions change quickly, despite its advantages. This review explores VADER's shortcomings in adjusting to the dynamic environment of online language.   
Research such as (Pollyanna et al., 2014) criticize VADER's static lexicon, which although large, might not be able to keep up with the quick changes in language on Twitter and other platforms. For every lexicon-based tool, there is an emerging set of memes, slang, and syntactic patterns.

VADER's adaptability is often increased by research addressing these issues by adding machine learning algorithms to it. For example, to integrate VADER with adaptive learning models, which can update the sentiment lexicon in real time based on contextual cues and emerging trends in social media text. VADER works well for static or historical datasets, but because online language is so dynamic, it needs to be improved when used for real-time sentiment analysis.

1. Semantic Analysis

The significance of social media platforms, such as Twitter, as reliable sources of up-to-the-minute information has been progressively acknowledged in studies analyzing content. However, it is still uncertain to what degree these platforms can be equated with traditional news media. In order to rectify this inadequacy, a substantial investigation carried out by

(Xin Zhao et al., 2011) scrutinized whether Twitter should be regarded exclusively as a more rapid channel for news updates or as a distinct and unique stream of information.

They utilized an unsupervised topic modeling technique by employing a Twitter-LDA model to examine a representative sample of Twitter data. Their approach entailed extracting subjects from Twitter and connect them with those addressed by the New York Times. This analysis primarily examined the categories and types of topics, as well as the correlation between the proportions of opinionated tweets, retweets, and their corresponding topics.   
  
Their research uncovered notable discrepancies and similarities in the subject matter between Twitter and the New York Times, providing valuable insights into the unique and common areas of information dissemination on these platforms. The study also highlighted the ability of Twitter to supplement traditional news media, which has implications for future applications in Information Retrieval (IR) and Data Mining (DM). This research is extremely relevant to the study as it highlights the significance of social media platforms in the modern information ecosystem. This study provides a foundation for further exploration into how these digital platforms can potentially improve or challenge the role of social media, particularly in the context of financial news.

The study conducted by (Blei M., Y. Ng and Jordan, 2003) emphasizes the ability of LDA to offer a clear depiction of document topics by means of topic probabilities. LDA is a generative probabilistic model that is used to describe collections of discrete data, specifically text corpora. The model can be defined as a hierarchical Bayesian model consisting of three levels. This model represents each item in a collection as a finite combination of topics, and each topic is represented as an infinite combination of probabilities related to the topic. Furthermore, they offer a thorough elucidation of efficient approximate inference techniques utilizing variational methods and an Expectation-Maximization (EM) algorithm to estimate empirical Bayes parameters. The efficacy of LDA is assessed by its application in document modeling, text classification, and collaborative filtering, where it is contrasted with other models such as a mixture of unigrams model and the probabilistic Latent Semantic Indexing (LSI) model. The study also utilizes a Bayesian parameter estimator to optimize the LDA parameters, which is essential for improving the performance of the model. This study offered valuable insights into the application of parameter tuning in research.

Coherence measurement in sets of statements has various applications in fields like text mining and information retrieval. A significant contribution to this field is the research conducted by (Roder, Both and Hinneburg, 2015), which specifically investigated the coherence of topics generated by topic models. Their research is centered around the widespread issue of the lack of interpretability assurances in the output of these models.The study developed a novel framework that allows for the creation of both existing word-based coherence measures and new ones by integrating fundamental components. This innovative approach facilitated a systematic exploration of the coherence measure space, utilizing all publicly available topic relevance data for evaluation. Their research demonstrates that novel combinations of these elements display superior performance in relation to their correlation with human evaluations of topic interpretability, surpassing current metrics.

The outcomes of their research enhance the reliability of topic models and also possess the potential to enhance the precision of text mining, refine search algorithms in information retrieval, and further optimize content relevance on the World Wide Web.

The study conducted by (Hamed et al., 2017) examines the applications and advancements of Latent Dirichlet Allocation (LDA), a technique that is highly relevant in the field. The researchers conducted a thorough examination of scholarly articles published from 2003 to 2016. Their primary goal was to employ LDA-based topic modeling to monitor the advancement of research, detect prevailing trends, and visually represent the intellectual structure of the field. The study's findings demonstrate the adaptability of LDA across various domains, such as medical sciences, software engineering, and political science. Latent Dirichlet Allocation (LDA) has been utilized in software engineering to analyze and categorize software based on its source code. In the field of medicine, it helps to identify concealed patterns in large datasets, thus assisting in the diagnosis of diseases and the planning of treatments. The political implications of this technique are especially fascinating. It has been employed to analyze large volumes of political literature, assisting in the detection of underlying themes and patterns that may not be apparent through traditional reading methods.  
  
  
  
This summary provides a thorough overview of a study that uses LDA-based topic modeling to advocate for the use of Twitter data as a tool to analyze the connection between Twitter activity and real-world situations, as well as to identify pertinent news. Afterwards, this news will be incorporated into the RoBERTa model using the RAG application, which I have already optimized.

1. Attention Mechanism and Pre-Trained Models

The article "Attention Is All You Need" by (Vaswani et al., 2017), introduces the Transformer model, which presents a ground breaking change in the way sequence transduction tasks are approached. This model eschews the previously dominant architectures of recurrent and convolutional neural networks and instead embraces a system that relies exclusively on an attention mechanism. This innovative approach has had a significant influence on subsequent research in the field of NLP. The main innovation of the Transformer lies in its self-attention mechanism, which allows for the concurrent processing of input sequences. This leads to a substantial enhancement in speed and efficiency compared to earlier models that depend on sequential data processing. The Transformer architecture consists of an encoder and decoder, each composed of several layers of multi-head self-attention and position-wise fully connected feed-forward networks. Positional encodings are incorporated into the design of the Transformer to accurately represent the order of input sequences. This mitigates the lack of repetition and guarantees the preservation of the sequence order. The multi-head attention mechanism allows the model to selectively concentrate on various segments of the sequence, giving it a versatile advantage in handling different types of information within a single model. The practical implications of the Transformer model are evident through its outstanding performance on machine translation tasks, where it achieved state-of-the-art results upon its introduction. It exhibited exceptional performance in comparison to existing models when translating from English to German and English to French, highlighting its effectiveness and efficiency in training. The influence of "Attention Is All You Need" extends beyond improving particular tasks. Attention mechanisms have brought about a substantial transformation in models, resulting in the emergence of more sophisticated models like BERT and GPT. These models are constructed using the framework of the Transformer and have enhanced the capabilities of NLP in diverse applications.   
  
  
  
BERT (Bidirectional Encoder Representations from Transformers) is a highly significant advancement that has resulted from this innovation. The BERT model is a groundbreaking advancement in NLP Processing as it employs a unique bidirectional training method, enabling it to proficiently handle various NLP tasks. The architecture of BERT, as proposed by (Devlin et al., 2018), is founded on the transformer model. The transformer model employs an attention mechanism to ascertain the relative significance of words in a sentence.The primary innovation of BERT is its pre-training process, which entails training on an extensive corpus of text for two specific tasks: masked language modeling (MLM) and next sentence prediction (NSP). Pre-training enhances the model's ability to understand the relationships between words and sentences, thereby increasing its effectiveness in tasks such as question answering, language inference, and others.BERT has gained significant recognition in the academic field for its outstanding performance on multiple NLP benchmarks, including SQuAD v1.1 (Stanford Question Answering Dataset), GLUE (General Language Understanding Evaluation), and SWAG (Situations With Adversarial Generations). The system's effectiveness arises from its profound bidirectional nature, which allows it to understand the context of a word by considering not only the preceding words but also its surroundings.The versatility of BERT has led to its application in diverse domains beyond basic NLP tasks. It has demonstrated its effectiveness in domains such as biomedical text analysis, sentiment analysis, and even in aiding epidemiological research during the COVID-19 crisis. Specialized variations of BERT, such as CovBERT, which is trained on scientific literature related to COVID-19, have been utilized for this purpose.  
  
  
  
Several models have modified and extended BERT's structure to improve or customize its functionalities. RoBERTa is an improved version of the BERT model developed by (Devlin et al., 2018) which stands for Robustly Optimized BERT Pretraining Approach. RoBERTa, developed by (Liu et al., 2019), makes adjustments to several important hyperparameters in BERT in order to further enhance its performance. This has been discussed in multiple studies and replication attempts.

The training methodology of RoBERTa diverges substantially from that of BERT in various aspects. At first, it removes the task of predicting the next sentence (NSP), a change that has been found to improve performance on different benchmarks. The model focuses solely on the masked language modeling (MLM) task, which entails predicting randomly masked tokens in the input data. This modification seeks to address the problem of insufficient training for BERT and proposes that extending the duration of pretraining could improve the performance of the model. RoBERTa implements a significant enhancement by utilizing dynamic masking instead of the static masking employed in BERT. Static masking is the act of permanently fixing the tokens that have been masked before the training begins. Nevertheless, this approach has the potential to limit the breadth of variation in the learning process. Dynamic masking refers to the process of generating a new masking pattern each time data is entered into the model. This approach improves the training process by providing a more complex and varied learning experience, which exposes the model to a wider range of learning situations. RoBERTa sets itself apart through its training regimen and utilization of datasets. The model is trained on a large collection of data called a corpus. This corpus includes datasets like OpenWebText, CC-NEWS, BookCorpus, Stories (a subset of Common Crawl), and English Wikipedia. In total, the corpus contains over 160GB of text. This extensive training is carried out on hardware configurations that incorporate multiple Nvidia V100 GPUs, enabling the processing of larger batches and longer training sequences. The practical consequences of these modifications are robust and dependable. RoBERTa not only achieves similar performance but frequently outperforms the original BERT model on well-known NLP benchmarks like GLUE, RACE, and SQuAD. This system is particularly acknowledged for its ability to handle long sequences and its resilience in dealing with complex language comprehension tasks. RoBERTa has emerged as a renowned model for diverse NLP applications, underscoring the importance of optimizing hyperparameters and conducting comprehensive pretraining to develop robust language models.

1. Large Language models and the Retrieve and Generate (RAG) Framework

The research conducted by (Arefeen, Debnath and Chakradhar, 2023) focuses on the fast growth of incorporating Large Language Models in various fields such as healthcare, education, and customer service, specifically for question-answering applications. Nevertheless, the implementation of LLMs, particularly in a context that is specific to a particular domain, presents considerable cost obstacles due to the high expenses associated with API calls necessary for handling substantial amounts of data. The study aims to tackle these challenges by introducing "LeanContext," a system specifically developed to enhance cost efficiency while preserving the effectiveness of domain-specific QA systems.  
  
LeanContext utilizes reinforcement learning to dynamically assess the level of context reduction required for specific queries, striking a balance between cost effectiveness and the need to retain sufficient context for accurate query responses. This approach is highly innovative because it deviates from summarization strategies that prioritize human understanding and instead focuses on creating summaries that are more beneficial for AI models. The emphasis is on preserving contextually relevant information that directly aids the question-answering process.  
  
The need for such a system is emphasized by the inherent limitations in the current LLM setups, specifically the restrictions on input prompt length and the significant expense associated with larger contexts. By intelligently reducing the context to the most pertinent information, LeanContext significantly reduces operational costs without a proportional loss in performance, which is evidenced by minimal drops in ROUGE scores a common metric for evaluating the quality of text summaries and machine-generated content.

TThis study expands on established methodologies such as chunk-based processing and document summarization. However, it goes a step further by incorporating these techniques into a machine learning framework that adapts to the specific requirements of the query. Introducing a reinforcement learning model to dynamically choose context that is relevant to the query is a significant advancement compared to static summarization methods. Static methods often either remove too much useful information or keep unnecessary data, both of which can negatively impact the performance of question-answering systems.

Comparatively, LeanContext offers a more refined approach to handling domain-specific data for QA tasks by not only ensuring that the cost is kept to a minimum but also that the quality of the responses is not compromised. The system’s ability to maintain context integrity while reducing unnecessary content could set a new standard for cost-effective LLM deployment in domain-specific applications.

Furthermore, the study's empirical findings demonstrate the efficacy of LeanContext in cost reduction while simultaneously upholding a high degree of precision. The system successfully achieves substantial cost reductions in API usage (up to 67.81%), while experiencing only a slight decrease in ROUGE-1 scores (ranging from 1.41% to 2.65%). These results are particularly promising, offering a path forward for small businesses and other entities that might otherwise be unable to afford the high costs associated with deploying cutting-edge AI technologies in a domain-specific context.

In its conclusion, this study contributes to the field of AI and machine learning by addressing a significant barrier to the widespread implementation of LLMs: their high cost. LeanContext enhances the input context to ensure excellent output quality, allowing the use of LLMs in settings with limited resources and promoting the implementation of advanced AI technologies in different industries.  
  
The recent study offers a thorough investigation into the process of optimizing pre-trained language models (PLMs) for domain-specific question answering (QA) tasks while dealing with limited resources. This research is crucial because it tackles the practical difficulties of implementing advanced NLP techniques in situations where there is a shortage of annotated data, but there is still a pressing need for high accuracy.  
  
The fundamental concept of the study revolves around the ineffectiveness of conventional sequential fine-tuning methods in situations where there is a scarcity of annotation resources. Alternatively, the research proposes a mixed dataset approach that combines domain-specific data with high-quality general data sourced from datasets such as SQuAD. This approach greatly improves the model's performance without requiring extensive additional annotations.

The research conducted by (Guo et al., 2024) contributes to the ongoing discussion in the NLP community on how to improve the cost-effectiveness of implementing language models in specific fields. The study rigorously assesses various fine-tuning strategies through comprehensive testing on multiple QA datasets. The study determines that a combined fine-tuning method, which integrates target-specific data with general QA data, achieves a better trade-off between performance enhancement and annotation cost, especially in settings with limited resources.  
  
An important discovery from this research is that using traditional methods such as masked language modeling (MLM) to pre-train on domain-specific collections of texts may not always improve performance and can actually worsen the performance of the model. This observation contradicts the commonly accepted belief in NLP fine-tuning and emphasizes the importance of developing strategies specifically tailored to a particular domain.

Moreover, the research offers a valuable resource for QA practitioners by providing a comprehensive comparison of 108 distinct fine-tuning strategies implemented on four varied datasets. This comprehensive analysis not only emphasizes the most effective methods but also outlines the constraints of current approaches, especially in terms of their ability to adapt to different budget levels.  
  
The study enhances the discourse on the suitability of large pre-trained models such as BERT and RoBERTa for domain-specific tasks, in terms of theoretical contribution. This text provides a more detailed explanation of how the effectiveness of question-answering systems can be improved by combining various fine-tuning techniques, such as aligning tasks and fine-tuning target data.  
  
This literature makes a significant contribution to the field by providing a detailed and practical method for improving the training of QA systems while working within limited financial resources. This is particularly important for small to medium-sized businesses and research institutions that want to benefit from NLP technology without having to spend a lot of money on data annotation.  
  
The results of this study have significant implications, offering possibilities for future research to explore more cost-effective and efficient methods for implementing language models in specific fields. This study represents a crucial advancement in making advanced NLP more attainable and useful for real-world applications, particularly in cases where data annotations are expensive and require a significant amount of time.  
  
Retrieval-Augmented Generation refers to the process of improving the ability of a model, such as Transformer, to generate content by including a retrieval component. This component extracts relevant information from a comprehensive external knowledge base and incorporates it into the generation process. The objective of this approach is to provide models with a broader range of specific and contextually relevant data, which could potentially improve the model's ability to produce more precise and contextually appropriate responses.

The study of (Xie et al., 2022) emphasizes the division in structured knowledge grounding (SKG) tasks across various domains, where each subset of tasks has evolved separately, frequently employing incompatible methodologies and datasets. The proposed framework, called UNIFIEDSKG, aims to consolidate these divergent approaches into a coherent structure. By implementing a standardized method of using RAG for all twenty-one different SKG tasks, a more organized and efficient approach to research and application in this field is enabled. The inclusion of RAG in this framework has demonstrated substantial improvements in the performance of the language model on different benchmarks. The findings from these implementations suggest that incorporating retrieval processes into generative models not only improves their factual accuracy but also enhances their efficiency in task-specific applications. For example, models that were adjusted and improved using this unified framework achieved the best results in tasks such as semantic parsing and natural language generation. This underscores the efficacy of RAG in augmenting the capabilities of the models. Furthermore, the document examines the adaptability of RAG in handling diverse data structures such as web tables, knowledge graphs, and databases, showcasing its capability to effectively process various sources of information. Adaptability is crucial for applications in domains that require accuracy and up-to-date information, such as medical diagnosis, legal counsel, and customer support. This study also explores the application of RAG in zero-shot and few-shot learning scenarios, where it allows models to achieve satisfactory performance even with limited task-specific training. The incorporation of this feature in RAG has the capacity to significantly reduce the time and resources required for the successful implementation of AI solutions in diverse domains. It opens up new possibilities for the broader adoption of RAG technologies, suggesting a future where AI systems can access and utilize human knowledge with greater efficiency and accuracy.

The research conducted by (Alawwad et al., 2024) centers on improving textbook question answering (TQA) through the integration of a large language model and Retrieval-Augmented Generation. TQA is an advanced artificial intelligence task that involves handling various types of data. Conventional language models often struggle to comprehend and grasp the context of lengthy textbook content. This has resulted in the creation of sophisticated solutions like Llama-2, a Language Model renowned for its exceptional performance in diverse NLP benchmarks. The paper presents a method to enhance the precision of TQA by incorporating RAG, which retrieves relevant contextual information from dispersed lessons. This technique tackles the "out-of-domain" challenge, which refers to the situation where vital information is distributed among various lessons. This approach offers comprehensive and meticulous responses by considering all pertinent data, regardless of its placement within the curriculum. The empirical results show significant improvements, with a 4.12% rise in accuracy seen in the validation set and a 9.84% rise in the test set for non-diagram multiple-choice questions. This study sets a new benchmark in artificial intelligence for education by demonstrating the successful application of Language Models and Retrieval-Augmented Generative models to tackle the intricate difficulties of educational content.

|  |  |  |  |
| --- | --- | --- | --- |
| Author(s) | Year | Applications | Data Analytics Techniques |
| Blei, D. M., Ng, A. Y., & Jordan, M. I. | 2003 | Generative probabilistic model for collections of discrete data such as text corpora. | EM, Bayesian parameter estimator |
| Zhao, Wayne X., Jiang Jing, Weng Jianshu, He Jing, Lim Ee-Peng, Yan Hongfei, Li Xiaoming | 2011 | Comparison of news and twitter data with LDA. | LDA, Topic Modeling |
| Eric Gilbert, C.J. Hutto | 2014 | Analyzing the Sentiment of Tweets using various sentiment analysis tools and presenting the VADER lexicon alongside a comparison to other methods. | Benchmarkings: LIWC, ANEW, the General Inquirer, SentiWordNet. ML Methods: Naive Bayes, Maximum Entropy, SVM, VADER-lexicon, F1-Score. |
| Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, Meeyoung Cha | 2014 | Sentiment Analysis Tools Comparison | Emoticons, LIWC (Linguistic Inquiry and Word Count), SentiStrength, SentiWordNet, SenticNet, SASA (SailAil Sentiment Analyzer), Happiness Index, PANAS-t (Positive Affect Negative Affect Scale adapted for Twitter). |
| Roder, Both and Hinneburg | 2015 | Enhancing the interpretability and reliability of topic models in text mining and IR. | LDA, Coherence score |
| Liu Y., Wang Y., Zhao Y., Li Z. | 2020 | Mental health disparities in the transgender community using social media sentiment analysis | Kappa Score, Bag-of-Words, TF-IDF, Naïve Bayes, Random Forest, SVM, Logistic Regression, K-Nearest Neighbour, CNN, LSTM |
| Arefeen, M. A.; Debnath, B.; Chakradhar, S. | 2023 | Domain-specific question-answering systems | Context Reduction (LeanContext), BERT, RoBERTa, ROUGE score |
| Kunpeng Guo, Dennis Diefenbach, Antoine Gourru, Christophe Gravier | 2024 | Fine-Tuning LLMs with limited data and low-cost. | 108 Different fine-tuning methodology compared. |
| Hessa Abdulrahman Alawwad, Areej Alhothali, Usman Naseem, Ali Alkhathlan, Amani Jamal | 2024 | TextBookQA is an educational dataset for question answering, containing questions suitable for Grade 6 students. | Llama2, RAG, TextBookQA |

# References

Agarwal, M., 2019. An Overview of Natural Language Processing. *International Journal for Research in Applied Science and Engineering Technology*, 7(5), pp.2811–2813.

Alawwad, H.A., Alhothali, A., Naseem, U., Alkhathlan, A. and Jamal, A., 2024. Enhancing Textbook Question Answering Task with Large Language Models and Retrieval Augmented Generation. arXiv:2402.05128v2. .

Arefeen, M.A., Debnath, B. and Chakradhar, S., 2023. LeanContext: Cost-Efficient Domain-Specific Question Answering Using LLMs. *arXiv:2309.00841*.

Arora, A.S. and Sanni, S.A., 2019. Ten Years of ‘Social Media Marketing’ Research in the Journal of Promotion Management: Research Synthesis, Emerging Themes, and New Directions. *Journal of Promotion Management*, pp.476–499.

Blei M., D., Y. Ng, A. and Jordan, M.I., 2003. Latent Dirichlet Allocation.

Chakraborty, U. and Bhat, S., 2018. The Effects of Credible Online Reviews on Brand Equity Dimensions and Its Consequence on Consumer Behavior. *J. Promot. Manag.*, pp.57–82.

Chin, W.W., Marcelin, B.L. and Newsted, P.R., 2003. A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study. *Inf. Syst. Res.*, pp.189–217.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*.

Dijkmans, C., Kerkhof, P. and Beukeboom, C.J., 2015. A stage to engage: Social media use and corporate reputation. *Tourism Management*, [online] pp.58–67. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0261517714001757?via%3Dihub>.

Djafarova, E. and Rushworth, C., 2017. Exploring the credibility of online celebrities’ Instagram profiles in influencing the purchase decisions of young female users. *Comput. Hum. Behav.*, pp.1–7.

Dodoo, N.A., 2018. Why Consumers Like Facebook Brands: The Role of Aspirational Brand Personality in Consumer Behavior. *J. Promot. Manag.*, pp.103–127.

Dwairi, R.A., Harb, Y. and Shehabat, I., 2020. The role of social platforms in establishing brand awareness for B2C e-commerce: empirical study. *International Journal of Business Information Systems*, 33(4), p.488.

Ferreira, M. and Zambaldi, F., 2019. The Mediating Role of Consumer Engagement with the Brand Community and Its Effect on Corporate Reputation. *International Journal on Media Management*, 21(1), pp.45–64.

Freberg, K., Graham, K., McGaughey, K. and Freberg, L.A., 2011. Who are the social media influencers? A study of public perceptions of personality. *Public Relat. Rev.*, (37), pp.90–92.

Guida, H. and Wilson, O., 2017. Social Identity Matters: Social Media and Brand Perceptions in the Fashion Apparel and Accessories Industries. *DOI: 10.4018/978-1-5225-2697-1.CH016*. [online] Available at: <https://api.semanticscholar.org/CorpusID:168704650>.

Guo, K., Diefenbach, D., Gourru, A. and Gravier, C., 2024. Fine-tuning Strategies for Domain Specific Question Answering under Low Annotation Budget Constraints. *arXiv:2401.09168v1*.

Hamed, J., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L., 2017. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *ArXiv*.

Helal, G. and Ozuem, W., 2021. Social Media and Social Identity in the Millennial Generation. In: *Research Anthology on Strategies for Using Social Media as a Service and Tool in Business*. IGI Global. pp.161–200.

Hutto, C. and Gilbert, E., 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: *Proceedings of the International AAAI*. Conference on Web and Social Media. pp.216–225.

Ismail, A.R., Nguyen, B. and Melewar, T.C., 2018. Impact of perceived social media marketing activities on brand and value consciousness: roles of usage, materialism and conspicuous consumption. *International Journal of Internet Marketing and Advertising*, 12(3), p.233.

Khan, M.A., Alhathal, F., Alam, S. and Minhaj, S.M., 2023. Importance of Social Networking Sites and Determining Its Impact on Brand Image and Online Shopping: An Empirical Study. *MDPI (Multidisciplinary Digital Publishing Institute)*. [online] Available at: <https://doi.org/10.3390/su15065129>.

Khan, M.T., Durrani, M., Ali, A., Inayat, I., Khalid, S. and Khan, K.H., 2016. Sentiment analysis and the complex natural language - complex adaptive systems modeling. *Springer Berlin Heidelberg*. [online] Available at: <https://casmodeling.springeropen.com/articles/10.1186/s40294-016-0016-9>.

Kumar, V., Bhaskaran, V., Mirchandan, R. and Shah, M., 2013. Practice prize winner-creating a measurable social media marketing strategy: Increasing the value and ROI of intangibles and tangibles for hokey pokey. pp.194–212.

L. V. Thang, Jodie Conduit, D. Q. H. Hai, and H. Nguyen, 2016. The impact of social media engagement on customer brand perception: The mediating effect of electronic word-of-mouth.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Facebook AI*.

Liu, Y., Wang, Y., Zhao, Y. and Li Zhixiang, 2020. TRANSGENDER COMMUNITY SENTIMENT ANALYSIS FROM SOCIAL MEDIA DATA: A NATURAL LANGUAGE PROCESSING APPROACH. *arXiv preprint arXiv:2010.13062*. [online] Available at: <https://arxiv.org/abs/2010.13062>.

Pollyanna, G., Araújo, M., Benevenuto, F. and Cha, M., 2014. Comparing and combining sentiment analysis methods. [online] Conference on Online Social Networks. . Available at: <https://api.semanticscholar.org/CorpusID:17347124>.

Ranco, G., Aleksovski, D., Caldarelli, G., Grcar, M. and Mozetic, I., 2015. The Effects of Twitter Sentiment on Stock Price Returns. *ArXiv*. [online] Available at: <https://doi.org/10.1371/journal.pone.0138441>.

Rizwan Ali Khadim, B. Zafar, and M. Younis, 2014. Social Media Communication and Consumer Brand Perceptions.

Roder, M., Both, A. and Hinneburg, A., 2015. Exploring the Space of Topic Coherence Measures. In: *WSDM ’15: Proceedings of the Eighth ACM*. International Conference on Web Search and Data Mining. pp.399–408.

Severyn, A. and Moschitti, A., 2015. Twitter Sentiment Analysis with Deep Convolutional Neural Networks. In: *Proceedings of the 38th International ACM SIGIR*. [online] Conference on Research and Development in Information Retrieval. Association for Computing Machinery. pp.959–962. Available at: <https://dl.acm.org/doi/abs/10.1145/2766462.2767830>.

Van Doorn, J., Lemon, K.N., Mittal, V., Nass, S., Pick, D., Pirner, P. and Verhoef, P.C., 2010. Customer engagement behavior: Theoretical foundations and research directions. *Journal of Service Research*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention Is All You Need. 31st Conference on Neural Information Processing Systems. .

Wirtz, J., Den Ambtman, A., Bloemer, J., Horvath, C., Ramaseshan, B., Van de Klundert, J. and Kandampully, J., 2013. Managing brands and customer engagement in online brand communities. *Journal of Service Management*.

Xie, T., Wu, C.H., Shi, P., Zhong, R., Scholak, T., Yasunaga, M., Wu, C.-S., Zhong, M., Yin, P., Wang, S.I., Zhong, V., Wang, B., Li, C., Boyle, C., Ni, A., Yao, Z., Radev, D., Xiong, C., Kong, L., Zhang, R., Smith, N.A., Zettlemoyer, L. and Yu, T., 2022. UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models. pp.602–631.

Xin Zhao, W., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X., 2011. Comparing Twitter and Traditional Media Using Topic Models. In: *In Advances in Information Retrieval*. European Conference, Dublin, Ireland. .