Artificial Intelligence and Social Computing, Vol. 163, 2025, 193–203
https://doi.org/10.54941/ahfe1006051

Enhancing Utility Customer Service and Compliance: An AI-Powered Approach to Call Analysis

Jonathan Presto[1] and Kar Wai Lee[2]

[1]Southern California Edison, 2131 Walnut Grove, Rosemead, CA 91770, USA
[2]Southern California Edison, 6090 N Irwindale Ave, Irwindale, CA 91702, USA

ABSTRACT

This study presents a framework for analyzing customer service call transcripts using a Large Language Model (LLM) and unsupervised machine learning. We employed BERTopic to identify core topics from summarized transcripts, refined through an iterative process against internal best practices. The LLM then generated detailed call reasons and agent responses, mapped to standardized tags via an embedding model for consistency. This framework, implemented on a scalable GCP architecture with robust security measures, allows for granular root cause analysis and identification of customer sentiment trends. Evaluation of the LLM demonstrated high recall rates for topic detection and accuracy in generating summaries and call reasons. This approach enables initiative-taking identification of customer needs, targeted agent coaching, and compliance risk mitigation, enhancing customer experience and operational efficiency.

Keywords: Large language models (LLM), Speech analytics, Bertopic, Prompt engineering, Customer service, Call center analytics, Topic modeling

INTRODUCTION

In today's data-driven world, customer service interactions are a goldmine of insights that can significantly enhance business operations and customer satisfaction. Analyzing customer service call transcripts is crucial for understanding customer needs, identifying pain points, and improving service quality. Traditional methods of analyzing these interactions, such as manual review or basic keyword extraction, are often limited in scope, scalability, and accuracy. They fail to capture the nuanced and dynamic nature of customer conversations.

This study leverages the power of Large Language Models (LLMs) and unsupervised machine learning to provide a comprehensive framework for topic discovery and sentiment analysis in customer service interactions. By utilizing advanced techniques like BERTopic for topic modeling and LLMs for text generation, we can automatically uncover hidden themes and patterns within vast collections of call transcripts. This approach not only enhances the granularity and accuracy of the analysis but also scales efficiently to manage large volumes of data.

The primary goal of this research was to develop a scalable framework, BERTopic and LLM-Augmented Speech Analytics (BLASA), for analyzing call transcripts, identifying core topics, and generating detailed call reasons and agent responses. This framework aims to enable businesses to proactively identify customer needs, provide targeted agent coaching, and mitigate compliance risks. It seeks to enhance the overall customer experience and operational efficiency by providing actionable insights derived from customer interactions.

This paper introduces a novel approach combining BERTopic and LLMs for topic discovery and text generation on a scalable Google Cloud Platform (GCP) architecture. Key contributions include a robust framework integrating BERTopic for topic modeling and LLMs for generating detailed call reasons and agent responses, designed for large-scale data processing in customer service contact centers. The scalable GCP architecture ensures efficient processing of vast call data, with robust security measures protecting sensitive customer information.

Initial evaluations demonstrate the framework's effectiveness, showing high recall rates for topic detection and accuracy in generated summaries and call reasons. By capturing customer interaction nuances, the framework also provides deeper insights, such as potential root causes and granular call reason tags with paired agent responses. These insights help businesses monitor call-handling quality, understand call drivers and customer sentiment, and take corrective actions to improve customer experience and operational efficiency.

RELATED WORK

Analyzing customer service interactions has long been a focus of past research in natural language processing (NLP). Traditional techniques often involve rule-based systems (Hirschberg and Manning, 2015) and basic keyword extraction methods (Turney, 2000). These approaches, while useful, are limited in their ability to capture the complexity and nuance of human conversations. For instance, rule-based systems can struggle with the variability in language use, and keyword extraction methods may miss context-specific meanings.

Recent advancements in LLMs, such as GPT-3 and BERT, have revolutionized the field of NLP by enabling more sophisticated text understanding and generation (Kumar and Singh, 2023). These models leverage deep learning techniques to process and generate human-like text, making them highly effective for a wide range of applications, including sentiment analysis, text summarization, and topic modeling (Kheiri and Karimi, 2023; Singh and Paridhi, 2023). Studies have shown that LLMs can significantly improve the accuracy and depth of insights derived from text data compared to traditional methods (Boitel et al., 2024; Alhijawi et al., 2024).

Topic Modeling With BERTopic

BERTopic is an unsupervised machine learning technique that has gained popularity for its ability to identify and visualize topics within large text

corpora. Unlike traditional topic modeling methods like Latent Dirichlet Allocation (LDA), BERTopic leverages BERT embeddings to capture semantic relationships between words, resulting in more coherent and meaningful topics (Mishra et al., 2024). BERTopic method was applied in various domains, but its use in analyzing customer service call transcripts is novel. This study aims to fill this gap by demonstrating the effectiveness of BERTopic in uncovering hidden themes and patterns in customer interactions.

However, one of the challenges with BERTopic is finding the optimal parameters to achieve the best results. BERTopic can frequently output raw topics that are overlapping or redundant. To address this, a hierarchical tree output can be useful to visualize and decide if certain topics should be merged or split (MaartenGr, 2024). This process, however, often relies on collaboration between model developers and domain experts to refine the final list of topics, ensuring minimal overlap and optimal separation.

Traditional customer service interaction analysis relies on manual review or basic keyword extraction, which is time-consuming and labor-intensive, limiting scalability and comprehensive insights. While customer-centric companies use advanced NLP techniques, the integration of LLMs and sophisticated topic modeling methods like BERTopic remains underutilized (Gana et al., 2024). BERTopic's output can guide AI prompt creation for downstream detection of cleaner topics independently, generating indicator variables for each topic. These variables can then be used as labels to measure the relevancy or accuracy of more granular call reason standardized tags derived from AI-generated call reason texts.

LLMs for Automated and Scalable Solution

BLASA offers a more automated and scalable solution for analyzing customer service call transcripts. By combining BERTopic for topic discovery with LLMs for text generation, we can provide deeper and more accurate insights into customer interactions. This framework not only reduces the need for manual intervention but also enhances the granularity and relevance of the analysis. The implementation on a scalable GCP architecture further ensures that the solution can manage large datasets efficiently, making it suitable for real-world applications in customer service.

To highlight the novelty and effectiveness of our approach, we compare it with existing methods in both academia and industry. Traditional NLP techniques and basic keyword extraction methods are contrasted with our framework, highlighting the improvements in accuracy, scalability, and depth of insights. Additionally, we discuss the limitations of current industry practices and how our solution addresses these challenges, providing a more comprehensive and automated approach to customer service analytics.

METHODOLOGY

The foundation of our framework is built on a comprehensive dataset of customer service call transcripts. These transcripts are collected from various customer service call types over a specified date range, ensuring a diverse and representative sample. The raw transcripts often contain a significant

amount of irrelevant content, such as greetings, small talk, and other non-essential dialogue. To address this, we employ a Large Language Model (LLM) to summarize these lengthy interactions, focusing on extracting the core keywords and phrases that are most relevant to the analysis.

The objective of the preprocessing steps is to remove irrelevant content, reduce noise and improve the quality of the data. Pre-processing includes text summarization and normalization, allowing for the data fed into the BERTopic model to be clean, relevant, and ready for topic discovery (Groot et al., 2022).

BLASA Framework Overview

Our framework systematically analyzes customer service call transcripts through a series of well-defined steps. Steps 1–7 is model development phase and steps 8–9 is model inference phase:

1. **Summarization:** The LLM generates summaries of the call transcripts, focusing on the most relevant information.
2. **Topic Discovery:** BERTopic algorithm processes summarized transcripts to identify core subtopics and themes.
3. **Iterative Refinement:** We update the collection of subtopics iteratively by comparing them against internal best practices and emerging trends.
4. **Topic Identification and Refinement:** Subject matter experts review the updated collection of subtopics to ensure alignment with business objectives.
5. **Topic Detection Prompt Creation:** Finalized subtopics can guide the creation of AI detection prompts to flag future call records for specific subtopics. These indicator variables also serve as labels to later validate the accuracy of more granular reason tags.
6. **Call Reason and Agent Response Generation:** After finalizing the collection of subtopics, the LLM generates detailed descriptions of call reasons and agent responses, capturing the sequence of events and nuances of each interaction.
7. **Mapping to Standardized Tags:** Each generated call reason text is assigned to a standardized tag by fine-tuning a sentence transformer to measure semantic similarity (Reimers and Gurevych, 2019).
8. **Tag Assignment:** A standardized tag is determined based on the highest similarity score, ensuring consistent categorization of similar phrases (e.g., 'high bill,' 'increased billing amount,' and 'high account balance' map to 'HIGH/INCREASED BILL CONCERNS'). We also defined a taxonomy to group tags into subcategories and further group subcategories into broader categories to support downstream reporting.
9. **Subtopic Classification:** The LLM detects one or multiple subtopics applicable to each call transcript. We group the subtopics under broader topics to also support downstream reporting.

Detecting subtopics and generating call reasons serve distinct yet complementary purposes in our BLASA framework (Fig. 1). Subtopics provide a broader overview of the main themes in the call dataset, which helps in understanding the drivers for incoming calls. This detection process

enriches the call dataset with indicator variables, one for each subtopic. These indicator variables, also serve as labels to validate the accuracy of the granular reason tags. This dual approach ensures both high-level and detailed insights, allowing for cross-verification of results to enhance reliability. Additionally, subtopics frame the context in which call reasons occur, providing a richer understanding of customer interactions.

For call reasons text generation, we instruct the language model to output up to twenty call reasons, with the first reason being the most probable root cause. We also ask the model to rationalize the reasons in sequential order to understand the potential sequence of events. This helps us explore whether one call reason could be the cause of another, like performing causal analysis. By combining both methods, we achieve a comprehensive analysis that captures emerging trends and specific customer concerns, improving the accuracy and actionability of our insights.

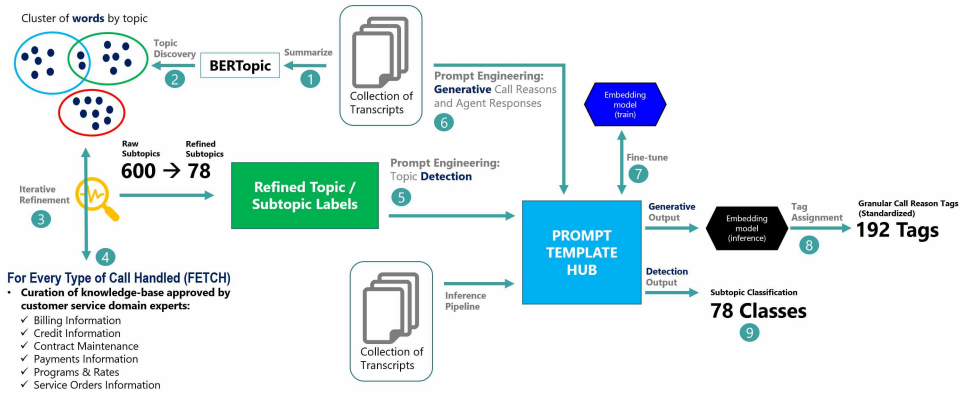

Figure 1: BLASA applies AI Prompts to detect subtopics and generate call reason text.

IMPLEMENTATION

Our framework (Fig. 2) leverages the robust and scalable infrastructure of Google Cloud Platform (GCP) and Snowflake to efficiently manage large volumes of data with high security and reliability. The process begins with data ingestion and preprocessing, where raw call transcripts are imported from Snowflake into GCP. Initial steps include cleaning, noise reduction, summarization using the LLM, and normalization.

Once pre-processed, the data is partitioned into smaller batches, each assigned to a dedicated worker node, significantly reducing processing time from days to hours. Task orchestration is managed within a GCP Workbench instance, which schedules and monitors Python-based scripts. Central to our processing is the Prompt Template Hub, a curated repository of prompts guiding the interaction with the LLM. Each transcript is embedded within a prompt, instructing the LLM to output its response in JSON format, which is then parsed and converted into a dataframe for analysis.

To ensure fault tolerance and data integrity, we implemented a function that periodically saves mini batches to a Google Cloud Storage (GCS) bucket,

safeguarding against data loss during processing. After initial processing, the data is enriched with standardized call reason tags using a fine-tuned sentence transformer, a specific type of embedding model, to map AI-generated reason text to a predefined taxonomy, ensuring consistency and facilitating downstream analysis. The enriched dataset is written back to Snowflake, undergoing further SQL transformations to generate reporting views.

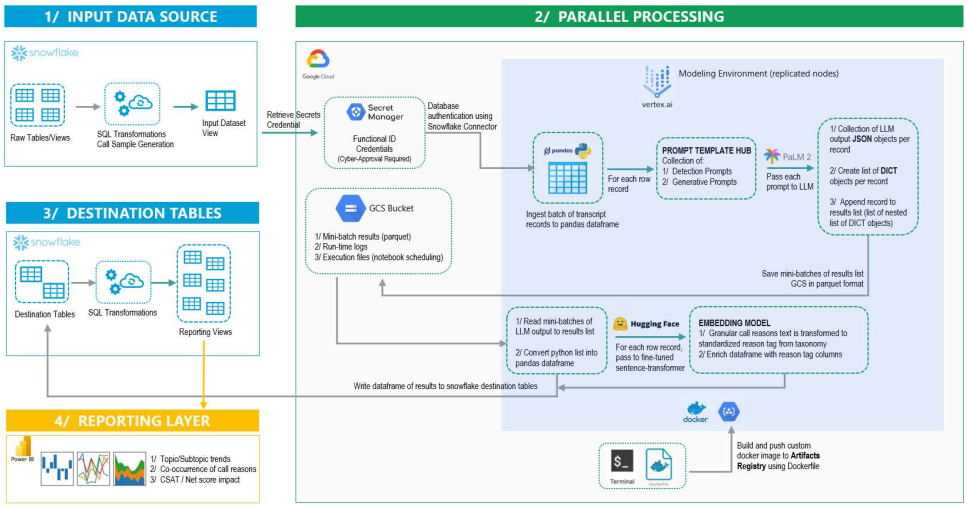

Figure 2: BLASA batch processing framework & technology stack.

Reporting views feed data to reporting dashboards, providing business users with intuitive visualizations and actionable insights. Security is ingrained in our infrastructure, using redacted transcripts to protect customer privacy and secrets management service for secure credential storage (Hetz et al., 2024). Daily executions rely on building and pushing custom virtual environments, ensuring reproducibility and meeting package dependency requirements.

By leveraging the scalability and security features of GCP, our implementation ensures efficient processing, robust fault tolerance, and secure handling of customer data. This architecture supports comprehensive analysis of call drivers and trends (Figs. 3, 4), as well as potential root cause analysis with net score impact that could drive business improvements. For instance, by identifying key call drivers, we can implement targeted strategies to enhance customer satisfaction and operational efficiency.

Moreover, this framework is platform-agnostic and can be implemented on other modern cloud compute platforms that support large-scale machine learning and efficient data processing of big data. This flexibility ensures that our solution can adapt to various environments and leverage the best available technologies.

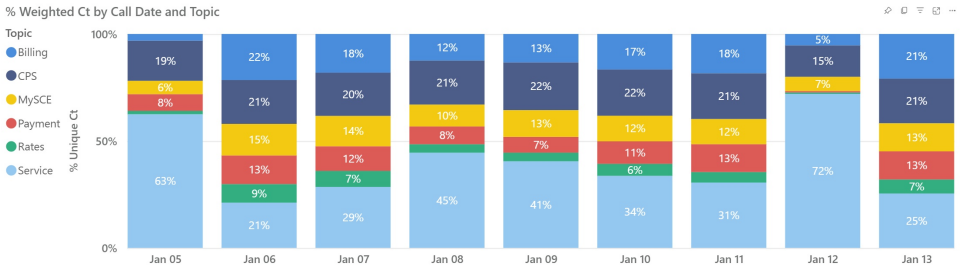

Figure 3: Call driver trends by topic.

Topic/Subtopics by Call Date

Topic	2025-01-05	2025-01-06	2025-01-07	2025-01-08	2025-01-09	2025-01-10	2025-01-11	2025-01-12	2025-01-13	Total
⊞ Billing	4%	66%	49%	24%	28%	39%	45%	7%	58%	47%
⊞ CPS	26%	63%	55%	41%	48%	51%	52%	20%	57%	54%
⊞ MySCE	9%	45%	38%	20%	27%	28%	29%	9%	36%	34%
⊞ Payment	11%	41%	31%	16%	16%	25%	32%	1%	36%	30%
⊞ Rates	2%	27%	20%	8%	9%	13%	12%	1%	18%	17%
⊟ Service	87%	65%	77%	88%	87%	79%	75%	98%	70%	76%
CROSSED_DEVICES		2%	2%	1%	1%	1%	1%	1%	2%	1%
DASR		0%	0%		0%	1%	0%		0%	0%
FOIE_OIE		2%	1%	0%	1%	1%	1%		2%	1%
GRID_RESILIENCY	2%	1%	2%	6%	8%	5%	4%	9%	2%	3%
ISO_ALERTS		0%		0%		0%	0%		0%	0%
METER	7%	11%	13%	5%	6%	7%	8%	6%	12%	10%
MOVE_IN	7%	29%	27%	13%	14%	14%	15%		20%	21%
MOVE_OUT		12%	11%	6%	7%	9%	9%	1%	13%	10%
NEW_ORDERS		7%	7%	2%	2%	5%	1%	1%	8%	5%
ORDER	48%	15%	24%	35%	23%	20%	28%	49%	20%	23%

Figure 4: Call driver trends deep dive by subtopic.

MODEL EVALUATION AND PERFORMANCE

Evaluating the performance of our model involves key metrics and methodologies to ensure accuracy, reliability, and overall effectiveness. First, we assess our confidence in the generated summaries. Next, we evaluate the precision and recall of subtopic assignment. Finally, we examine the relevancy of granular reason tags using a heuristic-based approach. Given the complexity and scale of the data, traditional manual review is impractical over a large validation dataset. Instead, we employ a combination of manual and semi-automated techniques to assess the quality of our models.

The first step in model evaluation was to measure the accuracy of AI-generated summaries against the original call transcripts using human reviewers. Due to the complexity of nuanced interactions between agents and customers, this manual evaluation was crucial to ensure the coherence, relevance, and completeness of the AI-generated text. Human reviewers assigned a confidence category to each generated output, which was then used to derive a confidence-driven accuracy score (Table 1, Table 2).

Table 1: Call summary evaluation guide.

Confidence Level	Guidelines
High	- Summaries accurately reflect the call content with minimal to no errors. - Key points and details are clearly captured. - The summary provides a coherent and comprehensive overview of the call. - Easily understandable by anyone unfamiliar with the original call.
Medium-High	- Summaries are mostly accurate with few minor errors (e.g., slight misinterpretations) but remain comprehensible. - Key points are mostly clear, with minor details occasionally missed. - The summary provides a good overview of the call, with minimal effort needed to understand.
Medium	- Summaries have noticeable errors (e.g., omissions, misinterpretations) that affect understanding. - Key points and details might be unclear in some sections. - The summary provides a partial overview of the call, requiring effort and context to understand the full meaning.
Low	- Summaries contain many errors (e.g., missing key points, incorrect information) that significantly hinder comprehension. - Key points and details are difficult to distinguish. - The summary provides a fragmented overview of the call, requiring significant effort and reference to the original call to understand.

Table 2: Evaluation results.

Confidence Level	CT	%
High	96	48 %
Medium-High	89	45 %
Medium	15	8 %
Low	0	0 %
TOTAL	200	100 %
Yes (High; Medium-High)	185	93 %
No (Medium; Low)	15	8 %
TOTAL	200	100 %

Upon reviewing two hundred randomly selected transcripts, the evaluation achieved an 92.5% accuracy rate for generated summaries. Notably, 48.0% of the summaries received a "High" confidence rating from human reviewers and 44.5% received a "Medium-High," highlighting the model's capability to produce coherent and relevant summaries that accurately reflect the call transcript content.

In the second step we measured how well the model assigns subtopics. Since there are seventy-eight subtopics, each is treated as an independent binary classification task, meaning the model must decide whether each subtopic applies to a given transcript. Two key metrics, precision and recall, measure model performance. Recall captures the count of correct subtopics the model identifies, ensuring comprehensive coverage. Precision reflects how

many assigned subtopics are correct, reducing false positives. A high recall ensures relevant subtopics are detected, while a high precision minimizes incorrect assignments.

Initially, we lacked labeled data for subtopics. Given the manual effort to read lengthy transcripts, we estimated recall using a sample of three hundred random transcripts and assigned subtopic labels, then compared them with the subtopic predictions. For estimating precision, we manually reviewed fifty predictions per subtopic then assigned TRUE or FALSE if we agreed with the prediction. Through iterative refinement, we improved subtopic detection prompts, achieving an overall recall rate of 85% and an overall precision rate of 89%.

The third step in the evaluation process was to assess the accuracy assigning a standardized call reason tag to the AI-generated reason text. Each AI-generated reason text was mapped to one of the 192 standardized reason tags. We then applied a heuristic-based approach that mapped each standardized tag to one or more subtopic. Given the complexity and substantial number of calls (10K ~ 13K daily calls), manually reviewing a large validation dataset would seem impractical; but a heuristic-based approach allowed us to evaluate model performance using a larger dataset.

To evaluate LLM model performance in a more scalable fashion, we propose a new metric called *relevancy-based accuracy*, or simply *relevancy* (Faggioli et al., 2023). Unlike traditional accuracy metrics that treat predictions as strictly correct or incorrect, relevancy accounts for the degree to which an assigned reason tag aligns with the intended meaning and context by using the subtopic flags that were also detected in the call. Of the 192 reason tags, ten examples illustrate this concept in Table 3. If we expand this heuristic-based approach to the rest of the other reason tags, then we can minimize manual review efforts while maintaining a balance between automation and quality control.

Table 3: Model evaluation using rule-based relevancy metric.

Example	Standardized Reason Tag	Relevancy	CT	Subtopic Flags Triggered
1	Financial. Hardship or Payment Affordability	98%	8734	CPS__DISCONNECT; CPS__LATE_FEE; CPS_ PAST_DUE; CPS IGP, CPS__CARE CPS_FERA CPS_IP.CPS__AMP: BLIING__ GENERAL
2	Care Program Discounts and Credits	91%	115	CPS__IQP; CPS__CARE; CPS__FERA; CPS__ PAST_DUE
3	High/Ncreased BLLL Concerns	91%	13134	BILUNG__GENERAL; BILUNG__HIGH _USAGE; BILUNG__HIGH_BLLL
4	Medical Baseline Program	88%	360	CPS__MBL: CPS__I 10 P

Continued

Table 3: Continued

Example	Standardized Reason Tag	Relevancy	CT	Subtopic Flags Triggered
5	Solar/Settlement Bill Inquiries, Credit and Rebates	86%	730	EILING__GENERAL; BILING___NEM_BLLL; EILLING ESTMATED BILL BILUNG__CORRECTION, BILLING__DISPUTE
6	Cross Metering	78%	188	SERVCE_-METER; SERMCE _CROSSED_DEVICES, SERVICE_ORDER
7	Tenant Move-IN/Move-OUT or Service Transfer	74%	4420	SERMCE_-MOVE INS, SERVICE_- MOVE_OUT; SERVICE_- NEW_ORDERS, SERVCE__TRANSFER _CONTRACT
8	Clean and Show	72%	309	SERMCE_-MOVE_IN: SERVMCE _-_MOVE_OUT; SERVCE__NEW_ORDERS, SERVCE__TRANSFER _CONTRACT
9	Move-IN/Move-OUT or Transfer/S Tartistop Service	70%	19855	SERUCE_-MOVE_IN, SERVICE__MOVE_OUT; SERVICE__NEW_ORDERS: SERVCE_- TRANSFER_CONTRACT
10	Heap Program	53%	34	CPS__HEAP, CPS__ICP; CPS__ESA

CONCLUSION

This study presents a scalable framework for analyzing customer service call transcripts using BERTopic and LLMs. The framework demonstrates high recall rates and accuracy in topic detection and summary generation, highlighting its effectiveness in capturing the nuances of customer interactions. By combining the strengths of BERTopic for topic modeling and LLMs for text generation, the framework provides a comprehensive solution for understanding and improving customer service operations.

Future research could explore the application of BLASA to other data sources, such as social media platforms, survey responses, healthcare records or legal documents. Expanding the framework to include these additional data sources would provide a more holistic view of customer interactions and sentiment across multiple channels, enabling businesses to gain deeper insights into customer behavior and preferences. Additionally, utilizing a Retrieval-Augmented Generation (RAG) framework can help minimize hallucination and improve the reliability of generated text by combining the strengths of retrieval-based and generation-based models, ensuring the text is grounded in trusted information (Gupta et al., 2024). Establishing robust evaluation metrics for LLMs is crucial to ensure their effectiveness and alignment with human values. Future research should prioritize standardized frameworks that address accuracy, fairness, and ethics. This will advance

customer service analytics, enhance AI applications, and enable businesses to fully leverage data for innovation and improved customer experiences.

REFERENCES

Alhijawi, B., Jarrar, R., AbuAlRub, A., & Bader, A. (2024). Deep learning detection method for large language models-generated scientific content. Neural Computing and Applications, 37, 91–104.

Boitel, E., Mohasseb, A., & Haig, E. (2024). A comparative analysis of GPT-3 and BERT models for text-based emotion recognition: Performance, efficiency, and robustness. Advances in Computational Intelligence Systems, 1453, 567–579.

Faggioli, G., Chiticariu, L., Neumann, M., & Oard, D. W. (2023). Perspectives on large language models for relevance judgment. arXiv preprint arXiv:2304.09161.

Gana, B., Leiva-Araos, A., Allende-Cid, H., & García, J. (2024). Leveraging LLMs for Efficient Topic Reviews. Applied Sciences, 14(17), 7675.

Groot, M., Aliannejadi, M., & Haas, M. R. (2022). Experiments on Generalizability of BERTopic on Multi-Domain Short Text. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP). doi: 10.18653/v1/2022.emnlp-main.70.

Gupta, S., Ranjan, R., & Singh, S. N. (2024). A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. arXiv preprint arXiv:2410.12837.

Hetz, G., Franzen, C., & Aliannejadi, M. (2024). Whisper-NER: Integrating Automatic Speech Recognition with Named Entity Recognition for Enhanced Privacy. Proceedings of the 2024 Conference on Privacy and Data Protection in AI, 112–125. doi: 10.18653/v1/P24-1125.

Hirschberg, J., & Manning, C. D. (2015). Advances in Natural Language Processing. Science, 349(6245), 261–266.

Kheiri, K., & Karimi, H. (2023). SentimentGPT: Exploiting GPT for advanced sentiment analysis and its departure from current machine learning. Data Science and Applications Lab, Utah State University.

Kumar, D., & Singh, S. (2023). Advancements in transformer architectures for large language models: From BERT to GPT-3 and beyond. International Research Journal of Modern Engineering and Technology, 55985.

MaartenGr. (2024). Hierarchical Topics in BERTopic. GitHub.

Mishra, M., Vishwakarma, S. K., Malviya, L., & Anjana, S. (2024). Temporal analysis of computational economics: A topic modeling approach. International Journal of Data Science and Analytics.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992. doi: 10.18653/v1/D19-1410.

Singh, U., & Paridhi. (2023). Comparing transformer architectures for sentiment analysis: A study of BERT, GPT, and T5. International Journal of Innovative Research in Technology, 167124.

Turney, P. D. (2000). Learning Algorithms for Key phrase Extraction. Information Retrieval, 2(4), 303–336.