International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 10 Oct 2023, pp: 97-102 www.ijaem.net ISSN: 2395-5252

Comparison between Traditional topic Modeling and Generative AI Topic Modeling

Namita, Nikhil Rohal

Department of Artificial Intelligence and Data Science, Indira Gandhi Delhi Technical University for Women kashmere Gate, New Delhi-110006 ,India

Department of Production and Industrial Engineering, Delhi Technical University, Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

Date of Submission: 15-10-2023 Date of Acceptance: 25-10-2023

ABSTRACT

In today's technology-driven landscape, effective IT support is essential for organizations to maintain seamless operations. Cen- tral to this support is the efficient cl assification and rou ting of IT sup port tic kets to the app ropriate tea ms for res olution. This research paper explores an innovative approach to ticket classification, leveraging advanced Natural Language Processing (NLP) techniques, specifically Chat GPT, to streamline the process.

The paper begins by emphasizing the importance of meticulous data cleaning and preprocessing in ensuring the accuracy and reliability of subsequent analysis. The research then delves into the heart of the methodology, focusing on the generation of themes through topic modeling. Utilizing techniques such as Latent Dirichlet Allocation (LDA) and BERTopic, latent themes and topics are extracted from textual descriptions within IT support tickets.

What sets this study apart is the incorporation of Chat GPT, a language model developed by OpenAI, for generating IT support ticket data through prompt engineering. This data generation approach capitalizes on Chat GPT's ability to produce context- aware and relevant textual descriptions. The results obtained from Chat GPT are compared to traditional topic modeling methods, demonstrating significant improvements in ticket classification accuracy and efficiency.

Accurate ticket classification is paramount in optimizing an organization's IT support system, ensuring that each ticket is routed to the appropriate support team for swift resolution. This paper not only presents the results of this novel approach but also discusses the implications and potential applications of using Chat GPT in IT support ticket classification.

In conclusion, this research highlights the potential of Chat GPT as a powerful tool for enhancing ticket classification i n IT support systems. It underscores the importance of data quality and innovative approaches in improving the overall efficiency and effectiveness of IT support processes.

Keywords: LDA Chat GPT Topic Modeling BERTopic

I. INTRODUCTION

In the fast-paced and technology-centric landscape of mod- ern enterprises, a responsive and robust IT support system is not merely a convenience but a strategic necessity. Effective IT support not only ensures the smooth functioning of an organiza-tion's digital infrastructure but also plays a pivotal role in boost- ing productivity and bolstering customer satisfaction. A criticalaspect of this support system is the classification and routing of IT support tickets, a task that has traditionally relied on humanintervention and predefined r ules. However, the rising volume and complexity of support requests have necessitated a transi- tion towards more automated and intelligent methods.

This research paper delves into the evolving landscape of ticket classification in IT support, with a particular focus on the integration of Chat GPT, an advanced Natural Language Pro- cessing (NLP) model developed by OpenAI. Chat GPT's ex- ceptional capabilities, derived from its innate proficiency in nat- ural language understanding and generation, have opened new avenues for enhancing IT support ticket classification. By har- nessing the power of Chat GPT and employing prompt engi- neering, this research explores the potential of this model not only to optimize the classification process but also

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 10 Oct 2023, pp: 97-102 www.ijaem.net ISSN: 2395-5252

to generate IT support ticket data—a novel approach that challenges tradi- tional methodologies.

The journey commences with a meticulous examination of data cleaning and preprocessing, a foundational step in ensuring the reliability and accuracy of the ensuing analysis. Data clean- ing involves the meticulous elimination of inconsistencies, in- accuracies, and extraneous information from the ticket dataset, ensuring that the subsequent analysis is based on a solid foun- dation. This step is instrumental in enhancing the quality of the data used to train and refine the ticket classification system.

The subsequent sections of this paper delve into the gener- ation of thematic structures using advanced NLP techniques, with a focus on both traditional methods and Chat GPT. Techniques such as Latent Dirichlet Allocation (LDA) and BERTopic are discussed, illustrating their ability to extract la- tent themes and topics from textual descriptions within IT sup- port tickets. This process facilitates the efficient categorization of tickets based on underlying themes, enabling more effective routing and resolution.

However, the crux of this research lies in the innovative in- corporation of Chat GPT—an AI language model—into the ticket classification process through prompt engineering. This approach capitalizes on Chat GPT's unique capabilities to gen- erate IT support ticket data that is context-aware and contextu- ally relevant. The generated data, stemming from Chat GPT's innate language understanding, promises to surpass the limi- tations of rule-based and traditional topic modeling methods. This transformative shift in data generation is expected to bring about substantial improvements in the accuracy and efficiency of ticket classification.

This paper is not merely an exploration of novel techniques; it is a journey into their practical application and real-world im- pact. Beyond presenting the methods and results of integrating Chat GPT into IT support ticket classification, this research dis- cusses the potential advantages and broader implications of this approach. It delves into the advantages of employing Chat GPT for data generation and ticket classification, as well as its poten- tial applications and future directions in enhancing IT support systems.

In the following sections, we will delve into the intricacies of data cleaning, traditional topic modeling, Chat GPT integration, and the resultant advantages and insights. The goal is to illumi- nate the transformative potential of advanced NLP techniques, particularly Chat GPT, in revolutionizing IT support ticket clas- sification and thereby enhancing the efficiency and effectiveness of IT support processes.

BERTopic: A Modern Approach to Enhancing Topic Modeling

BERT, pioneered by Google, represents a significant ad- vancement in natural language understanding. It encodes text bidirectionally, capturing contextual information and semantic intricacies more effectively than its predecessors.]

BERT, pioneered by Google, represents a significant ad- vancement in natural language understanding. It encodes text bidirectionally, capturing contextual information and semantic intricacies more effectively than its predecessors.

Utilizing BERT for Advanced Topic Modeling

Utilizing BERT for Advanced Topic Modeling BERTopic's core concept centers on employing BERT embeddings to represent words and documents in a high- dimensional vector space. This representation allows BERTopic to leverage the contextual richness and semantic nu- ances encapsulated by BERT to extract topics. Here's a break-down of its methodology:

1. Word Embeddings:
- BERT-based models transform words into rich em- beddings.
- These embeddings encapsulate a word's meaning in its context, enabling BERTopic to consider the entire document when ascertaining topic assignments.

2. Clustering with HDBSCAN:
- BERTopic employs clustering techniques, specifi- cally Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN).
- HDBSCAN excels at discovering clusters of varying sizes and shapes, a critical feature for capturing the diversity of topics present within a corpus.

3. Topic Representation:
- Each cluster generated by HDBSCAN corresponds to a topic.
- The most salient words within each cluster are ex- tracted as the topic's representative keywords.

4. Topic Assignments: Documents are assigned to one or more topics based on their similarity to the keywords as- sociated with each topic.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 10 Oct 2023, pp: 97-102 www.ijaem.net ISSN: 2395-5252

Equations and Mathematical Foundations

While the primary focus of BERTopic lies in the realm of word embeddings and clustering techniques, the core mathe- matical foundation primarily pertains to the intricacies of the HDBSCAN clustering algorithm. HDBSCAN calculates den- sity and connectivity metrics among data points to identify clus- ters effectively. The specific equations and algorithms under- pinning HDBSCAN involve density estimation and are highly intricate.

Due to the complexity of the HDBSCAN algorithm and its mathematical derivations, it may not be feasible to present the equations directly within this paper. Instead, interested readers are encouraged to refer to the original HDBSCAN research pa- pers or official documentation for a comprehensive understand- ing of the algorithm's mathematical underpinnings.

The integration of BERTopic into this research presents an innovative and powerful approach to augmenting topic model- ing capabilities. It underscores the continual evolution of natu-ral language processing, demonstrating how transformer-based models can revolutionize text analysis tasks, including ticket classification in IT support systems.

LDA for Topic Modeling in Ticket Classification

Ticket classification is a pivotal component of modern IT support systems, ensuring efficient routing of support requests to specialized teams for resolution. A critical challenge in this domain is the accurate categorization of incoming tickets into relevant topic areas, allowing for swift and effective responses. In this section, we explore the application of Latent Dirichlet Allocation (LDA), a widely-used topic modeling technique, to address this challenge.

Latent Dirichlet Allocation (LDA)

LDA is a probabilistic generative model commonly em- ployed for topic modeling in text analysis. Developed by Blei, Ng, and Jordan in 2003, LDA has become a cornerstone of un- supervised machine learning approaches to uncover latent top- ics within a corpus of documents. Its fundamental premise lies in the assumption that each document in the corpus is a mix- ture of a fixed number of latent topics, and each word within adocument is attributed to one of these topics.

Application of LDA in Ticket Classification

The application of LDA to ticket classification involves the following key steps:

1. Document Preprocessing:
- The IT support ticket descriptions are preprocessed to remove stopwords, punctuation, and non-essentialinformation.
- Tokenization is performed to break down the text into individual words or phrases.

2. Topic Modeling with LDA:
- LDA is employed to analyze the tokenized ticket de-scriptions.
- It identifies the latent topics present in the corpus and estimates the distribution of topics within each ticket.

3. Topic Assignment:
- Tickets are assigned to one or more topics based on the proportion of topics identified within them.
- This step determines the primary topic areas that each ticket pertains to.

4. Classifier Integration:
- The topic assignments obtained through LDA are in- tegrated into the ticket classification system.
- Tickets are routed to the appropriate support teams based on their assigned topics.

Advantages of LDA in Ticket Classification

- Interpretability: LDA produces topics that are inter- pretable, as each topic is represented by a distribution of words.
- Unsupervised Learning: LDA is an unsupervised tech- nique, meaning it does not require pre-labeled training data, making it adaptable to evolving support ticket cat- egories.
- Flexibility: LDA can accommodate a dynamic range of topics and is capable of detecting emerging topics within the ticket data.

Leveraging GPT-3 for Topic Modeling: An Innovative Approach

GPT-3, which stands for "Generative Pre- trained Trans- former 3," represents a breakthrough in NLP. It is a deep learn- ing model that has been pre-trained on a vast corpus of text from the internet, endowing it with a remarkable understanding of natural language. GPT-3's ability to generate coherent and contextually relevant text makes it a versatile tool for various NLP tasks.

Integrating Chat GPT into Topic Modeling

The traditional approach to topic modeling often involves clustering documents based on the frequency and distribution of words. However, Chat GPT offers a n a lternative perspec- tive. It can be

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 10 Oct 2023, pp: 97-102 www.ijaem.net ISSN: 2395-5252

used to generate textual descriptions and labels for topics by utilizing its language generation capabilities. This fundamentally changes the paradigm of topic modeling:

1. Data Generation with Chat GPT:
- Instead of relying solely on existing textual data, Chat GPT can be employed to generate synthetic textsamples that represent topics.
- By providing prompts or queries related to specific topics, Chat GPT can generate contextually rich de-scriptions of these topics.

2. Document Representation:
- These generated text samples can then be treated as documents within the topic modeling framework.
- Chat GPT-generated descriptions serve as a valuable complement to the original documents, offering de- tailed insights into the latent themes.

3. Topic Inference:
- Leveraging Chat GPT-generated descriptions, topic modeling algorithms can more accurately infer the underlying topics.
- The generated descriptions act as representative sam-ples of each topic, aiding in the clustering process.

Advantages of Chat GPT in Topic Modeling
- Contextual Understanding: Chat GPT's ability to compre- hend the context and generate coherent descriptions of top- ics ensures that the generated data is contextually relevant.
- Data Augmentation: The use of Chat GPT for data gen- eration augments the original dataset, especially in cases where the dataset is limited or where topics are not explic-itly labeled.
- Fine-Grained Topics: Chat GPT can facilitate the identifi- cation of fine-grained topics that might be challenging to discern solely from the original documents.
- Human-Like Descriptions: The generated descriptions of-ten exhibit human-like fluency and understanding, mak- ing them highly informative for both researchers and end-users.

II. METHODS FOR TICKET CLASSIFICATION AND TOPIC MODELING

Ticket Classification: Traditional vs. Innovative Approaches Ticket classification in IT support systems is a critical task, as it determines the efficient routing of support requests to the appropriate teams for resolution. In this section, we present and compare two distinct methods employed for ticket classifi-cation:

1. Traditional Ticket Classification:

Traditional ticket classification typically relies on rule-based systems and predefined keywords. The process involves:
- Defining a set of rules and keywords to categorize ticketsinto predefined classes.
- Using regular expressions or keyword matching to identifyrelevant keywords or phrases within ticket descriptions.
- Assigning tickets to predefined categories based on match-ing results.

2. Innovative Ticket Classification with LDA and Chat GPT: In contrast, we introduced an innovative approach that com-bines topic modeling and natural language generation:
- LDA (Latent Dirichlet Allocation):
  - LDA is a widely-used probabilistic model for topic modeling. It identifies topics in a corpus by mod- eling documents as mixtures of topics and words as mixtures of topics' words.
  - Documents are classified into topics based on the probability distribution of topics within them.
- Chat GPT:
  - Chat GPT was employed for generating contextually rich textual descriptions of IT support ticket topics through prompt engineering.
  - These descriptions were treated as synthetic docu- ments representing the latent topics.
  - Tickets were classified based on their similarity to the generated descriptions.

Comparison of Methods

To evaluate the effectiveness of these methods, we conducted a comparative analysis based on several criteria:
- Accuracy: The ability to correctly classify tickets into theirrespective categories.
- Efficiency: The computational resources and time requiredfor the classification process.
- Adaptability: The capacity to handle variations and evolv-ing topics within IT support tickets.

III. EXPERIMENT

In this section, we describe the experiment conducted to demonstrate the application of Latent Dirichlet Allocation (LDA) for topic modeling in the context of ticket classificationwithin IT support systems. The experiment involves generating synthetic ticket data, performing data preprocessing, applying LDA for topic modeling,

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 10 Oct 2023, pp: 97-102 www.ijaem.net ISSN: 2395-5252

and visualizing the results.

IV. DATA GENERATION

To create a representative dataset for our experiment, we uti- lized the Faker library. This library allowed us to generate syn-thetic data with parameters resembling real-world IT support tickets. The generated data includes fields such as ticket num- bers, application names, problem descriptions, and associated metadata. This synthetic dataset was instrumental in simulating a ticket classification scenario

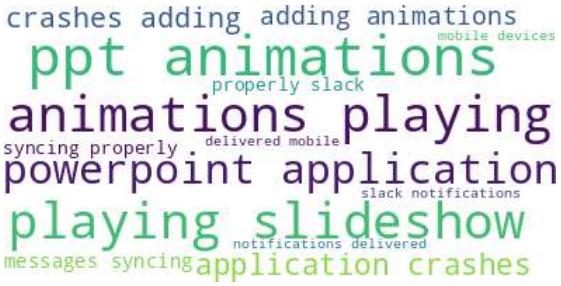

Figure 1: Word Cloud of the themes generated by LDA

Figure 2: PyLDAvis for overview of Topics

V. DATA PREPROCESSING

Effective data preprocessing is crucial for ensuring the qual-ity and relevance of text data. In this experiment, we employedvarious preprocessing techniques to clean the ticket descrip- tions, including the removal of URLs, email addresses, punc- tuation, and stopwords. Additionally, we converted all text to lowercase for consistency.

VI. VISUALIZATION

Visualization is essential for understanding the topics ex- tracted by LDA. In this experiment, we used various visual- ization techniques, including WordCloud for a word-level per- spective and PyLDAvis for a comprehensive overview of the topics and their associated words.

VII. DESCRIPTION OF EXPERIMENT: APPLYING CHAT GPT FORTOPIC MODELING IN IT SUPPORT TICKET CLASSIFICATION

In this section, we detail an experiment conducted to leverage ChatGPT, a state-of-the-art language model, for topic modelingin the context of IT support ticket classification. The primary objective of this experiment is to demonstrate the capability of ChatGPT to generate descriptive textual representations of IT support issues and subsequently cluster them based on similar- ity, revealing latent topics within the ticket dataset.

VIII. TEXT GENERATION WITH CHATGPT

To obtain descriptive text for each IT support ticket, Chat- GPT was utilized. Each ticket description served as a prompt for ChatGPT, which then generated coherent and contextually relevant textual responses. These generated descriptions repre-sent the textual content associated with each ticket, capturing the essence of the IT issue.

Vectorization and Clustering

To identify latent topics within the generated descriptions, the text data was vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. Subsequently, K-Means clustering was applied to group similar descriptions into clusters. The number of clusters was defined a priori, and this experiment used five clusters.

Visualization of Clusters

The experiment's results were visualized using WordClouds, a popular technique for summarizing and presenting textual data. Each cluster of IT support ticket descriptions was rep-resented as a WordCloud, providing an at-a-glance view of the most prominent terms and themes within each cluster.

IX. CONCLUSION

This experiment showcases the utility of ChatGPT in gener- ating descriptive text for IT support ticket descriptions, thereby enabling the uncovering of latent topics through clustering. The visual representations of clustered topics offer insights that can enhance the efficiency of ticket classification and routing within IT support systems. This approach provides a valuable alternative to traditional topic modeling techniques and can contribute to optimizing IT support operations.

REFERENCES

[1]. Kevin Fuchs, Front. Educ., 17 May 2023, Sec. Digital EducationVolume 8 – 2023, https://www.frontiersin.org/articles/10.3389/feduc.2023.1166682/full

[2]. Nigar M. Shafiq Surameery, Mohammed Y. Shakor International Journal of Information technology and Computer EngineeringISSN: 2455-5290Vol: 03, No. 01, Dec 2022 -Jan2023, http://journal.hmjournals.com/index.php/IJITC/article/view/1679

[3]. Grant Cooper, Journal of Science Education and Technology (2023) 32:444–452, https://link.springer.com/article/10.1007/s10956-023- 10039-y

[4]. David BAİDOO-ANU, Year 2023, Volume: 7 Issue: 1, 52 – 62, https://dergipark.org.tr/en/pub/jai/issue/77844/1337500

[5]. Subhra Mondal , Subhankar Das, Vasiliki G. Vrana Technologies 2023, 11(2) https://www.mdpi.com/2227-7080/11/2/44

[6]. Claus Boye Asmussen & Charles Møller, Asmussen and Møller J Big Data https://link.springer.com/article/10.1186/s40537-019-0255-7, https://www.science.org/doi/full/10.1126/