Universidade Estadual de Campinas
Instituto de Computação

Patrick Anderson Matias de Araújo

Topic Modeling Influence on Sentiment Analysis from User-Generated Product Reviews

Influência da Modelagem de Tópicos na Análise de Sentimentos a Partir de Avaliações de Produtos Geradas por Usuários

CAMPINAS
2024

Patrick Anderson Matias de Araújo

Topic Modeling Influence on Sentiment Analysis from User-Generated Product Reviews

Influência da Modelagem de Tópicos na Análise de Sentimentos a Partir de Avaliações de Produtos Geradas por Usuários

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. Julio Cesar dos Reis
Co-supervisor/Coorientador: Prof. Dr. Marcelo da Silva Reis

Este trabalho corresponde à versão final da Dissertação defendida por Patrick Anderson Matias de Araújo e orientada pelo Prof. Dr. Julio Cesar dos Reis.

CAMPINAS
2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Araújo, Patrick Anderson Matias de, 1994-
Ar15t Topic modeling influence on sentiment analysis from user-generated product reviews / Patrick Anderson Matias de Araújo. – Campinas, SP : [s.n.], 2024.

Orientador(es): Julio Cesar dos Reis.
Coorientador(es): Marcelo da Silva Reis.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Computação.

1. Aprendizado de máquina. 2. Processamento de linguagem natural (Computação). 3. Análise de sentimentos. 4. Inteligência competitiva (Administração). 5. Feedback. I. Reis, Julio Cesar dos, 1987-. II. Reis, Marcelo da Silva, 1979-. III. Universidade Estadual de Campinas (UNICAMP). Instituto de Computação. IV. Título.

Informações complementares

Título em outro idioma: Influência da modelagem de tópicos na análise de sentimentos a partir de avaliações de produtos geradas por usuários
Palavras-chave em inglês:
Machine learning
Natural language processing (Computer science)
Sentiment analysis
Business intelligence
Feedback
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Julio Cesar dos Reis [Orientador]
Helena de Almeida Maia
Thiago Henrique Silva
Data de defesa: 14-11-2024
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-9951-134X
- Currículo Lattes do autor: http://lattes.cnpq.br/5407186554644090

Universidade Estadual de Campinas
Instituto de Computação

INSTITUTO DE
COMPUTAÇÃO

Patrick Anderson Matias de Araújo

Topic Modeling Influence on Sentiment Analysis from User-Generated Product Reviews

Influência da Modelagem de Tópicos na Análise de Sentimentos a Partir de Avaliações de Produtos Geradas por Usuários

Banca Examinadora:

- Prof. Dr. Julio Cesar dos Reis
  Instituição de Computação / UNICAMP
- Dra. Helena de Almeida Maia
  Instituição de Computação / UNICAMP
- Prof. Dr. Thiago Henrique Silva
  Departamento Acadêmico de Informática / UTFPR

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 14 de novembro de 2024

Dedicatória

Este trabalho é dedicado primeiramente a duas pessoas fundamentais em minha vida: minha mãe, Luciana Matias de Barros, e minha irmã, Ananda Liz Matias de Araújo, cujo amor e apoio são inestimáveis. Em seguida, dedico ao meu padrasto, Oscar Yukiyoshi Ikeda, e à minha família, cujo suporte e presença são igualmente valorizados. Agradeço a todos por sua constante inspiração e incentivo, que tornaram este projeto possível. Não posso deixar de expressar minha gratidão à comunidade do Tocantins, em especial a Rita de Cássia Guimarães Melo e Paula Guimarães Dangelo, cuja ajuda foi crucial nesta trajetória.

O inconsciente é,
em seu fundo,
estruturado,
tramado,
encadeado,
tecido de linguagem.

(Jacques-Marie Émile Lacan)

Agradecimentos

Gostaria de expressar minha profunda gratidão a todos os membros da equipe do H.IAAC Meta 7 (IA para Marketing), com quem tive o privilégio de pertencer e colaborar. Em especial, agradeço ao graduando Gabriel Kenzo Kakimoto, aos mestrandos Débora Dias Panicachi e Sadeeq Olalekan Bello, e ao doutorando Fillipe dos Santos Silva.

Agradeço também aos doutores: Professor Doutor Eric David Cohen, Professor Doutor Júlio César dos Reis, Professor Doutor Marcelo da Silva Reis e Doutor Seyed Jamal Haddadi. Um agradecimento especial ao meu orientador, Professor Doutor Júlio César dos Reis, e ao meu coorientador, Professor Doutor Marcelo da Silva Reis.

No âmbito do H.IAAC, gostaria de destacar minha gratidão pela amizade dos amigos Arthur Hendricks Mendes de Oliveira e Juan David Nieto.

Agradeço a oportunidade de ter pertencido ao H.IAAC, que financiou minha pesquisa e proporcionou uma experiência enriquecedora. O Instituto de Computação da Universidade Estadual de Campinas foi o local ideal para meu mestrado, oferecendo uma experiência única que não poderia ser encontrada em nenhum outro lugar.

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado Arquitetura Cognitiva (Fase 3), DOU 01245.003479/2024-10.

Abstract

In today's data-driven economy, understanding customer feedback is crucial for businesses to improve their products and services. User reviews, representing a modern form of word-of-mouth marketing, offer valuable insights but are often vast and challenging to interpret manually. Advanced natural language processing (NLP) techniques provide a way to automatically extract meaningful insights from data, transforming how companies leverage customer input to drive strategic decisions. In this context, addressing topic identification and sentiment analysis in texts is crucial. However, the complexity of integrating topic modeling and sentiment analysis as two essential NLP techniques presents challenges, especially when applied to large-scale datasets with diversified linguistic features and informal languages, such as sarcasm, slang, and multilingual comments. This M.Sc. dissertation investigates the integration of topic modeling and sentiment analysis to enhance the interpretation of customer feedback. Focusing on user reviews from three major platforms – Amazon, Netflix, and Spotify – collected via the Google Play Store, our study leverages advanced natural language processing techniques to extract key information from user comments. In particular, we evaluate two approaches: 1) small fine-tuned NLP models such as BERT, T5, and BERTopic; 2) pre-trained large transformer models like Meta's Llama 3 8B and Mixtral 8x7B. Our investigation assesses how topic modeling affects sentiment classification across multi-class settings, examining its metrics (3 classes vs. 5 classes). Our research highlights the effectiveness of different NLP models, which give businesses a deeper understanding of customer behavior and enable data-driven strategic decisions. This M.Sc. dissertation contributes by demonstrating the relevance of combining topic modeling with sentiment analysis, advancing the application of AI in business intelligence.

Keywords: Machine Learning, Natural Language Processing, Sentiment Analysis, Competitive Intelligence, Feedback

Resumo

Na atual economia baseada em dados, entender o *feedback* do cliente é crucial para que as empresas melhorem seus produtos e serviços. As avaliações de usuários, que representam uma forma moderna de *marketing* boca a boca, oferecem *insights* valiosos, mas geralmente são vastas e desafiadoras de interpretar manualmente. Técnicas avançadas de processamento de linguagem natural (PLN) fornecem uma maneira de extrair automaticamente *insights* significativos dos dados, transformando a maneira como as empresas alavancam a entrada do cliente para conduzir decisões estratégicas. Nesse contexto, abordar a identificação de tópicos e a análise de sentimentos em textos é crucial. No entanto, a complexidade de integrar a modelagem de tópicos e a análise de sentimentos como duas técnicas essenciais de PNL apresenta desafios, especialmente quando aplicadas a conjuntos de dados em larga escala com características linguísticas diversificadas e linguagens informais, como sarcasmo, gíria e comentários multilíngues. Esta dissertação de mestrado investiga a integração da modelagem de tópicos e da análise de sentimentos para aprimorar a interpretação de críticas de clientes. Com foco nas avaliações de usuários de três plataformas principais – *Amazon*, *Netflix* e *Spotify* – coletadas por meio da *Google Play Store*, nosso estudo alavanca técnicas avançadas de processamento de linguagem natural para extrair informações importantes dos comentários dos usuários. Em particular, avaliamos duas abordagens: 1) usando modelos menores de PNL e ajuste fino, como *BERT*, *T5* e *BERTopic*; 2) usando modelos de transformadores pré-treinados grandes, como *Llama* 3 8B da *Meta* e *Mixtral* 8x7B. Nossa investigação avalia como a modelagem de tópicos afeta a classificação de sentimentos em configurações de várias classes, examinando suas métricas (3 classes vs. 5 classes). A pesquisa destaca a eficácia de diferentes modelos de PNL, que fornecem às empresas entendimentos mais profundos sobre o comportamento do cliente e permitem decisões estratégicas baseadas em dados. Esta dissertação de mestrado contribui ao demonstrar a relevância de combinar modelagem de tópicos com análise de sentimentos, avançando a aplicação de IA em inteligência de negócios.

Palavras-chave: Aprendizado de Máquina, Processamento de Linguagem Natural, Análise de Sentimentos, Inteligência Competitiva, *Feedback*

List of Figures

List of Tables

List of Abbreviations

ABSA Aspect-Based Sentiment Analysis

AI Artificial intelligence

API Application Programming Interface

ATC Automatic Topic Consolidation

BERT Bidirecional Encoder Representations from Transformers

c-TF-IDF class-based Term Frequency-Inverse Document Frequency

CNN Convolutional Neural Network

CPU Central Processing Unit

CSV Comma-Separated Values

GloVe Global Vectors

GPT Generative Pre-Trained Transformer

GPU Graphics Processing Unit

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

HIMT Hierarchical Interactive Multimodal Transformer

KGNMF Knowledge-guided Non-negative Matrix Factorization for Better Short Text Topic Mining

LDA Latent Dirichlet Allocation

LGPD Lei Geral de Proteção de Dados Pessoais

Llama Large Language Model Meta AI

LLM Large Language Model

LM Language Model

LS Label-Segmented

LSTM Long Short-Term Memory

MF-CNN-BiLSTM Multistage Feature Extraction using Convolutional Neural Network and Bidirectional Long Short-Term Memory

ML Machine Learning

MLP Multilayer Perceptron

M.Sc. Master of Science

NHS National Health Service

NLP Natural Language Processing

NLTK Natural Language Toolkit

NMF Non-negative Matrix Factorization

NPMI Normalized Pointwise Mutual Information

PMI Pointwise Mutual Information

Recod Reasoning for Complex Data

RNN Recurrent Neural Network

SO-CAL Semantic Orientation Calculator

T5 Text-to-Text Transfer Transformer

TF-IDF Term Frequency-Inverse Document Frequency

UCA Unified Corpus Analysis

UMAP Uniform Manifold Approximation and Projection

UTG Unrestricted Topic Generation

VADER Valence Aware Dictionary and Sentiment Reasoner

WOM Word of Mouth

Contents

Chapter 1

Introduction

1.1 Context and Motivation

Erik Erikson, the psychologist who introduced the concept of “identity crisis”, formulated the stages of psychosocial development. He emphasizes the significance of social interactions and developing a sense of belonging throughout one’s lifespan [83]. Social interactions play a crucial role in Word of Mouth (WOM) marketing, one of the most potent forms of promotion. People trust recommendations from friends, family, and peers more than traditional advertising. Successful marketing campaigns often aim to create buzz and conversations around their products or services within relevant social circles.

To remain competitive within a market and globalized economy, companies must enhance their products and services through informed decisions regarding development and marketing strategies [35]. Traditionally, customer experiences were shared through WOM. Today, we have more formal ways to capture this valuable information, such as customer feedback/reviews and social media comments. These insights significantly influence how consumers perceive brands and make purchasing decisions. Customer feedback, including reviews and social media comments, has become increasingly important for providing valuable observations. Customer feedback and reviews are methods for customers to give opinions on products or services, with the respondent being either identified or anonymized. Feedback is input actively solicited from customers through surveys or forms, while reviews are unsolicited evaluations left on marketplaces and review platforms, often including star ratings and written comments. When comments explicitly include a numerical rating, they can be considered graded reviews or feedback. Additionally, customers may leave more general remarks on company websites and social media pages. Whether identified or anonymous, feedback, reviews, and comments are valuable sources of customer voice and insight for businesses.

As consumer demands evolve, companies are under constant pressure to remain ahead. They can gain an edge by analyzing and promptly responding to customer feedback. This is crucial in industrial scenarios like car manufacturing because of the complex and costly development process. This ability to make informed decisions about product development and marketing strategies can be a significant strategic advantage. Therefore, businesses must continuously strive to improve and change their products and services to meet the changing needs of their diverse customer base.

Mobile devices have significantly changed the way customers give their feedback. With the rise of mobile technology, customers can now quickly provide feedback on the go, in real-time, and through multiple channels [24]. This has increased the frequency and volume of customer feedback, as customers can share their opinions and experiences anytime, anywhere. Mobile devices also enable customers to provide feedback through various formats, such as text, voice, images, and videos, giving businesses more affluent and detailed feedback.

1.2 Problem Characterization

The proliferation of digital feedback on e-commerce and entertainment platforms, representing a modern form of WOM, presents unique challenges. Many reviews can be found across various online sources, including but not limited to restaurant and hospitality websites, apps/games stores [92], and e-commerce platforms.

The enormous volume of reviews available poses a significant challenge to process and analyze them manually. These reviews can be written with different writing styles and cultural nuances, making extracting meaningful insights difficult. Nuances such as sarcasm, irony, and other complex language features [1, 42] make proper analysis challenging.

Nowadays, businesses can extract actionable insights from vast customer reviews across various platforms. Artificial intelligence (AI)[1] plays a crucial role in analyzing customer feedback across various industries, allowing companies to gain valuable information on which to base their decision-making processes.

Integrating Natural Language Processing (NLP) techniques like sentiment analysis and topic modeling can significantly address the challenges of interpreting customer feedback by uncovering underlying emotions and thematic patterns. NLP models are computational tools to do several tasks, they are also used to analyze and interpret customer feedback, specifically for tasks like sentiment analysis and topic modeling. These include Bidirecional Encoder Representations from Transformers (BERT) [16], BERTopic [22], Text-to-Text Transfer Transformer (T5) [57], Large Language Model Meta AI (Llama) [75], Mixtral [28], among others.

Sentiment analysis is a process to classify customer opinions to determine the emotional tone [53]. This task becomes complex when dealing with large volumes of unstructured text that may include ambiguous or conflicting sentiments. For instance, customers may express positive feelings through words but assign a low rating, or they may mix different emotions in one review [47]. Moreover, language features like irony, sarcasm, and cultural nuances make accurate sentiment detection difficult, especially when these subtleties are expressed in multiple ways across various reviews.

Topic modeling automatically identifies themes or subjects in a set of documents by analyzing patterns in word usage and co-occurrence [27]. A topic is defined as a set of related words that frequently occur together in a corpus, representing an underlying theme or subject. This is important because reviews often contain multiple intertwined topics,

[1] Artificial Intelligence is a field in Computer Science that concentrates on developing computer systems that can execute tasks requiring human intelligence intervention [62].

which can overlap with emotional sentiment. For example, a customer might mention both product quality and customer service in a single review, expressing different sentiments for each. This could give actionable insights to determine the correct sentiment, but it can present computational challenges that need to be addressed to maximize the insights gained. Topic granularity refers to the level of detail or specificity in the topics identified by topic modeling techniques. When analyzing a large corpus of text, these algorithms can generate topics at varying levels of granularity. Topics can be broad and general (low granularity), or more specific and detailed (high granularity) [48].

The challenge in analyzing customer reviews lies in the inherent complexity of text data. The diversity in vocabulary, tone, and writing style makes it difficult to extract clear insights using conventional methods. Sentiment analysis must detect subtle emotional cues, while topic modeling must handle overlapping themes within unstructured text. Combining these tasks is computationally demanding, requiring advanced NLP models to manage large datasets and extract both sentiment and thematic information.

Without effective integration of these techniques, businesses risk misinterpreting customer feedback, leading to flawed decision-making. Therefore, the integration of sentiment analysis and topic modeling is essential, though complex, in NLP.

1.3 Research Objectives and Questions

This Master of Science (M.Sc.) Dissertation aims to investigate whether the integration of topic modeling with sentiment analysis can enhance the effectiveness of sentiment classification across large volumes of customer feedback in English, thus enabling businesses to gain valuable insights. The specific objectives are as follows:

OBJ 1 **Investigate the Impact of Topic Modeling on Sentiment Analysis Metrics**: This study assesses how the integration of topic modeling techniques affects the effectiveness of sentiment analysis. The goal is to examine various metrics and determine if topic modeling improves the correctness of sentiment classification, either in a standard manner or with greater granularity[2].

OBJ 2 **Evaluate the effectiveness of NLP Models for Topic-Based Sentiment Analysis**: This objective focuses on comparing the precision of different NLP models, including small fine-tuned models and pre-trained large language models, when used in conjunction with topic modeling across diverse datasets. We assess how model effectiveness varies based on classification granularity.

OBJ 3 **Examine the Effect of Topic Granularity and Labeling on Sentiment Analysis**: This objective seeks to explore how the level of topic granularity and different topic modeling approaches impact sentiment analysis quality in different labeling

[2] Granularity refers to the level of detail or specificity in the analysis. In the context of sentiment analysis, greater granularity means examining sentiment classification at a more nuanced level, rather than the traditional categories of ‘negative,’ ‘neutral,’ or ‘positive.’ For example, a more granular approach might include classes like ‘very negative,’ ‘negative,’ ‘neutral,’ ‘positive,’ and ‘very positive.’ This approach is often referred to as multi-class classification, as it involves more than two classes.

classifications. This helps to identify the most effective strategies for handling large volumes of customer feedback.

Our research focuses on a critical and central question: how does integrating topic modeling affect sentiment analysis outcomes in analyzing customer reviews? Several specific and critical research questions arise in this context:

RQ 1 **Impact of Topic Modeling Techniques and Approaches on Sentiment Analysis Quality**: What is the impact of integrating topic modeling techniques on sentiment analysis metrics, and how do different topic modeling approaches influence the effectiveness across various datasets?

RQ 2 **Comparison of Small Fine-Tuned NLP Models and Pre-Trained Large Transformers with and without Topic Modeling information**: How do small fine-tuned NLP models compare with pre-trained large language models in sentiment classification tasks, both with and without the application of topic modeling?

RQ 3 **Effect of Topic and Sentiment Granularity on Sentiment Analysis**: How does topic granularity, ranging from high to low, affect sentiment analysis, and how does sentiment classification differ when employing three-class versus five-class sentiment approaches?

1.4 Significance and Justification

There is a lack of understanding regarding the specific factors that influence users to leave a good review or a bad one [69]. The need for sophisticated analytics to interpret complex customer feedback at scale has never been more pressing in our current world scenario.

Machine Learning (ML) algorithms can analyze customer feedback to identify patterns and provide actionable knowledge to decision-makers and marketing teams. For example, online marketplaces like Amazon and eBay[3] rely heavily on customer feedback to inform their product development strategies and improve the customer experience.

In critical industries such as healthcare and finance, customer feedback can play a crucial role in improving the quality and safety of services. The healthcare sector[4] uses patient feedback to assess and pinpoint opportunities for enhancing healthcare delivery. Similarly, financial institutions can utilize customer complaints and suggestions to enhance their security measures and customer service[5].

[3] Amazon: Using Big Data to understand customers. **Forbes**. Available at: <`https://www.forbes.com/sites/bernardmarr/2019/12/09/the-10-best-examples-of-how-companies-use-artificial-intelligence-in-practice`>. Accessed on June 18, 2024.

[4] Friends and Family Test. **National Health Service (NHS) England**. Available at: <`https://www.england.nhs.uk/fft/`>. Accessed on June 18, 2024.

[5] Twitter and LinkedIn became a tool for bank customers to have complaints resolved (*Twitter, LinkedIn e Reclame Aqui viram ferramenta para clientes de bancos terem queixas resolvidas*). **Valor Investe**. Available at: <`https://valorinveste.globo.com/produtos/servicos-financeiros/noticia/2023/04/05/twitter-linkedin-e-reclame-aqui-viram-ferramenta-para-clientes-de-bancos-terem-queixas-resolvidas.ghtml`>. Accessed on June 19, 2024.

1.5 Developed Approach and Originality

This research investigates how different topic modeling techniques affect sentiment analysis in multi-class settings, using app reviews from platforms like Amazon, Netflix, and Spotify. Our research focuses on two main approaches. The first approach involves small fine-tuned NLP models, such as BERT [16], T5 [57], and BERTopic [22]. The second approach includes pre-trained large language models like Meta's Llama 3 8B [75] and Mixtral 8x7B [28]. By comparing these approaches, our research aims to reveal their respective impacts on sentiment analysis quality for distinct datasets.

The first approach evaluated is called Small Fine-Tuned NLP Models and Techniques, where topic modeling with BERTopic can be approached by either segmenting the data or treating it as a whole, with the option to apply topic reduction techniques or leave the results unfiltered [48]. For sentiment analysis, BERT [16] and T5 [54, 57] models are fine-tuned to predict sentiment classes, both with and without topic modeling information to assess its impact on classification accuracy.

Our evaluation concerning the Pre-Trained Large Language Models approach employs Meta Llama 3 8B and Mixtral 8x7B to extract topics from user reviews. Initially, topics are generated for each user comment. These topics are then clustered and further refined to ensure clarity and relevance. This process allows the models to identify and consolidate meaningful themes, providing a rich representation of the topics in user feedback.

For sentiment analysis, the Pre-Trained Large Language Models are evaluated through different approaches that deal with how models learn from data. These methods investigate the models' ability to predict sentiment with minimal fine-tuning, assessing their adaptability and effectiveness in multi-class sentiment scenarios.

This research addresses a critical gap in the current scientific literature by examining how different configurations of topic modeling and sentiment analysis techniques influence product review classification. While topic modeling and sentiment analysis have been extensively studied [38], our study uniquely explores the interplay between various configurations of these techniques. We originally investigated how different topic modeling setups affect sentiment analysis models' quality metrics in multi-class scenarios. The quality of sentiment classification is evaluated using metrics such as F1-score. By analyzing these interactions, we provide novel insights into the optimal configurations for both topic modeling and sentiment analysis, aiming to improve the granularity and accuracy of sentiment classification.

Ultimately, this study determines whether integrating topic modeling with sentiment analysis enhances the precision and depth of sentiment identification from real-world user reviews. By comparing the Small Fine-Tuned NLP Models and Pre-Trained Large Language Models, we evaluate how these configurations improve sentiment analysis and give businesses more profound insights into customer feedback. These insights may support more informed strategic decisions and targeted product improvements.

By focusing on integrating topic modeling and sentiment analysis, this M.Sc. dissertation advances the broader field of NLP, demonstrating how thematic context can transform customer-centric business intelligence.

1.6 Organization of the Dissertation

The remaining of this M.Sc. dissertation is organized as follows:

Chapter 2 introduces key concepts essential to this research, including machine learning, NLP, and Large Language Model (LLM), with a focus on transformer architecture. It covers topic modeling techniques like BERTopic, as well as sentiment analysis methods. Ethical considerations, particularly data privacy, are also addressed, providing a solid foundation for understanding the technical and real-world implications of these technologies.

Chapter 3 reviews key research in topic modeling and sentiment analysis. It highlights the evolution of some techniques used from traditional to modern approaches, emphasizing their complementary use in analyzing both themes and sentiment in textual data. The integration of these methods is explored across various domains. The chapter also outlines how this research contributes to the literature.

Chapter 4 originally presents the framework designed and developed in this research for analyzing customer reviews using a combination of topic modeling and sentiment analysis techniques. It outlines the two main approaches. Then, the collected data undergoes preprocessing to filter out non-relevant or empty reviews, ensuring high-quality input. Evaluation metrics are used to compare the outcomes of the approaches, providing insights into the effectiveness of each model configuration in extracting meaningful sentiment and topics from user feedback.

Chapter 5 presents our experimental evaluation concerning the first approach developed. We detail the findings regarding the small fine-tuned NLP models and techniques employed.

Chapter 6 reports on our experimental evaluation to assess the use of pre-trained large language models. We discuss how this approach addresses the challenges of generating marketing-directed datasets for multi-platform applications.

Chapter 7 finalizes the research, focusing on the comparison of small fine-tuned NLP models and pre-trained large language models in sentiment analysis tasks. It discusses the importance of model selection and task complexity in optimizing results. Additionally, it addresses the study's limitations, including computational challenges, and suggests future research directions, such as exploring additional models, datasets, and advanced topic modeling techniques. The chapter provides a holistic understanding of the methodologies used and their implications for the field of NLP.

Chapter 2

Fundamental Concepts

This Chapter presents core aspects concerning Languade Models, offering a view of their architecture, underlying principles, and the mechanisms that enable their functionalities.

The Chapter aims to serve both newcomers and those familiar with the domain. It provides a structured overview that connects theoretical concepts with real-world applications and innovations in the field.

Section 2.1 introduces foundational concepts in ML and NLP, which are crucial for the following sections. Following this, Section 2.2 on Language Models examines the evolution of these models. In Section 2.3, the discussion shifts to topic modeling as a powerful NLP tool for discovering latent themes within large datasets. Finally, Section 2.4 introduces sentiment analysis, an NLP application focused on identifying emotions within text.

2.1 Machine Learning and Natural Language Processing

Machine Learning (ML) involves training models to make predictions or decisions without being explicitly programmed [37]. This is achieved through models and algorithms that learn from data and make predictions. There are three common learning approaches [85]: supervised learning, where the model is trained on labeled data to predict outputs from inputs; unsupervised learning, where the model identifies patterns in unlabeled data; and reinforcement learning lets machines learn by trial and error through rewards and penalties, making them useful for robotics, games, and navigation. Deep Learning [86], a subset of ML, uses neural networks with multiple layers to analyze data and is crucial for complex tasks like image and speech recognition, as well as natural language understanding.

NLP focuses on enabling interactions between computers and humans through language [21]. It involves techniques like tokenization, stemming, and lemmatization to preprocess text for machine understanding. Statistical Language Models, such as n-grams (contiguous sequence of n items from a given sample of text or speech), provide probabilities for word sequences and are useful for tasks like speech recognition and spell correction [30]. More recently, contextual embeddings like *Word2Vec*, Global Vectors (GloVe), and transformer-based models like BERT [16] and Generative Pre-Trained Transformer (GPT), have advanced NLP by capturing deeper levels of language context.

2.2 Language Models

In the field of NLP, language models serve as essential tools that enable applications requiring nuanced language understanding, interpretation, and generation. These models are fundamental to tasks such as sentiment analysis and topic modeling that this study aims to explore, and other text-driven insights. By learning patterns within language data, language models can make informed predictions about text, enhancing applications that need context-sensitive language processing.

Language models have evolved significantly, transitioning from basic statistical methods to advanced neural network architectures [87]. This evolution has improved how models handle complex language dynamics.

Early language models relied on statistical approaches, where simpler methods, such as n-grams, calculated the probability of words in a sequence. The conditional probability of the next word given the previous words is the core equation of the model. This is described as a likelihood of a word occurring given the prior words in a sequence, which is given by a simplified conditional probability: $P(w_n|w_{n-1}, w_{n-2}, \ldots, w_1)$, where w_n is the next word in the sequence, and w_{n-1}, w_{n-2}, and so on represent the preceding words [82].

Cross-entropy is a commonly used objective function in language modeling [81] that measures the effectiveness of a language model in predicting the next word in a sequence and the difference between the expected and actual distribution of words. Cross-entropy can be formalized as: $H(P, Q) = \sum_{x \in X} p(x) \log q(x)$, where $H(P, Q)$ is the cross-entropy between the true probability distribution p and the predicted probability distribution q, x represents the possible outcomes (in this case, words) in the distribution, $p(x)$ is the true probability of the result x, and $q(x)$ is the predicted probability of the outcome x [81].

While effective for shorter phrases, these models struggled with complex sentence structures and long-range dependencies. With the advent of the Recurrent Neural Network (RNN), language models gained the ability to handle longer text sequences and more contextual understanding further enhancing language modeling capabilities. However, RNNs are limited in scalability and often encounter challenges with long-range dependencies, as they process sequences in a step-by-step manner [40].

Introduced in 2017, the transformer architecture fundamentally changed language modeling by utilizing self-attention mechanisms that replace traditional sequence-aligned recurrence, enhancing parallelization and handling of long sequences [77]. This approach allows models to focus on the most relevant parts of an input sequence, regardless of its length, enabling them to grasp complex language patterns more efficiently. Transformers form the basis of current LLMs, such as BERT, T5, and GPT, which offer significant advancements in processing and generating human-like text. BERT processes text bidirectionally, meaning it considers context from both the left and right of a word in a sentence, enabling it to capture nuanced meanings and relationships [16]. While T5 is a versatile transformer model that frames all NLP tasks as a text-to-text problem, meaning both inputs and outputs are treated as text strings [57].

LLMs are advanced language models with billions of parameters, capable of processing and generating human-like text by leveraging transformer architectures. These models are distinguished by their ability to capture complex patterns in language data, enabling several tasks with high accuracy and contextual depth. The functioning of LLMs involves layers of multi-head attention and feed-forward networks within an encoder-decoder framework [77]. This structure enables them to efficiently process and generate text by focusing on relevant parts of the input data. Positional encodings and normalization layers further refine the model's output, helping it to manage long-range dependencies[10]. They have significantly impacted various fields by providing advanced capabilities in text generation, machine translation, and more. They promise innovations in multimodal learning and synthetic data generation [10].

The adoption of transformers has enabled models to perform various tasks central to this study. By capturing deeper context, these models can better interpret thematic nuances and sentiment within customer reviews, making them particularly useful for insights in customer-centric applications [77].

The use of language models raises ethical concerns, such as data privacy, potential biases in model outputs, and the risk of misinformation [10]. As language models become increasingly integrated into AI applications, future research aims to improve their ethical alignment, factual accuracy, and efficiency.

2.3 Topic Modeling

Topic modeling is a fundamental NLP technique that uncovers themes or topics within an extensive collection of documents. This method is invaluable for task classification, content recommendation, and information retrieval, as it helps reveal hidden structures within extensive text corpora [27].

At its core, topic modeling is an unsupervised learning technique that clusters words into topics based on their co-occurrence patterns within a dataset. Unlike traditional classification tasks, topic modeling does not require labeled data, making it especially useful for exploratory analysis where predefined categories are unknown.

One of the most widely used algorithms for topic modeling is Latent Dirichlet Allocation (LDA) [32], which represents each document as a mixture of topics and each topic as a word distribution. LDA assumes that documents are generated through a probabilistic process, where topics generate words with specific probabilities. The algorithm then works backward to infer the most likely topics that could have generated the observed documents. LDA operates through a few key steps:

1. **Initialization**: It begins by randomly assigning topics to each word in the corpus;
2. **Iteration**: The algorithm iteratively updates the topic assignments based on the likelihood of words belonging to topics, given the current topic assignments of other words;
3. **Convergence**: This process continues until the model stabilizes, resulting in a set of topics and corresponding word distributions that best explain the observed data.

In addition to LDA, other popular topic modeling approaches include Non-negative Matrix Factorization (NMF) and BERT-based topic models [56]. NMF works by factorizing the document-term matrix into two lower-dimensional matrices — one representing topics and the other representing documents — while BERT-based models leverage pre-trained contextual embeddings to enhance the quality of the generated topics.

An advanced topic modeling technique that has gained significant traction is BERTopic. BERTopic utilizes the powerful capabilities of BERT to generate meaningful topics from a collection of documents [22]. It combines several advanced techniques to achieve this:

1. **Document Embeddings**: BERTopic employs pretrained BERT models to create contextual embeddings for each document. These embeddings capture the semantic meaning of the text, going beyond mere word frequency;

2. **Dimensionality Reduction**: Given the high dimensionality of BERT embeddings, BERTopic uses Uniform Manifold Approximation and Projection (UMAP) to reduce this dimensionality. This process condenses the information into a 2D or 3D space, retaining the most significant features;

3. **Clustering**: To group similar document embeddings, BERTopic uses Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). This clustering technique organizes documents into clusters, each representing potential topics;

4. **Topic Representation**: Once clusters are formed, BERTopic interprets them as topics. It uses class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) to extract the most representative words for each topic. This step ensures that the topic descriptions remain dense and interpretable while preserving key terms;

5. **Dynamic Topic Modeling**: BERTopic also supports dynamic topic modeling, allowing it to track how topics evolve when applied to temporal data;

6. **Visualization**: Finally, BERTopic offers visualization options to explore the topics and their relationships, making understanding and analyzing the resulting topics easier.

A key factor behind the effectiveness of BERTopic is its utilization of c-TF-IDF for topic representation, a variant of TF-IDF. Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure used to assess the significance of a word within a document relative to a collection or corpus [22]. This measure increases with the frequency of a word in a document but decreases with its frequency across the entire corpus. The TF-IDF score for a term t in a document d within a collection of documents D is given by:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D). \tag{2.1}$$

It consists of two primary components:

1. **Term Frequency (TF)**: This indicates how often a term appears in a document. The underlying assumption is that a term that appears more frequently in a document is likely more relevant to its content:

$$\text{TF}(t, d) = \frac{\text{Count}(t, d)}{\text{Total terms in } d}. \tag{2.2}$$

2. **Inverse Document Frequency (IDF)**: This measures the rarity of a word across the entire corpus. Words that are frequent in many documents, such as 'the' or 'and,' have a low IDF score, as they provide little insight into the document's specific content. Conversely, terms that appear in fewer documents have a higher IDF score, highlighting their significance in distinguishing those documents:

$$\text{IDF}(t, D) = \log \frac{\text{Total number of documents in } D}{1 + \text{Number of documents containing } t}. \tag{2.3}$$

Combining TF and IDF, the TF-IDF score identifies terms that are particularly relevant to a specific document while not being overly common throughout the corpus. Thus, the TF-IDF score is given by:

$$\text{TF-IDF}(t, d, D) = \left(\frac{\text{Count}(t, d)}{\text{Total terms in } d} \right) \times \log \frac{\text{Total number of documents in } D}{1 + \text{Number of documents containing } t}. \tag{2.4}$$

In BERTopic, the c-TF-IDF modification refines this approach to focus on terms characteristic of specific clusters or topics, ensuring that the most representative terms for each topic are accurately identified. The c-TF-IDF score for a term t in a category/class c across all categories C in the corpus is given by:

$$\text{c-TF-IDF}(t, c, C) = \left(\frac{\text{Count}(t, c)}{\text{Total terms in } c} \right) \times \log \frac{\text{Total number of classes in } C}{1 + \text{Number of classes containing } t}. \tag{2.5}$$

Topic modeling, including advanced methods like BERTopic, has found applications across various domains. In marketing, it helps understand customer feedback by identifying recurring themes in reviews and social media posts. Academics use topic modeling to categorize and analyze large publications, revealing trends and patterns in research areas. In legal and regulatory fields, topic modeling assists in organizing and summarizing vast volumes of legal documents, facilitating more efficient information retrieval [90].

Despite its strengths, topic modeling faces challenges such as the interpretability of topics and the determination of the optimal number of topics. The results can also be sensitive to the choice of hyperparameters and preprocessing steps, such as stopword removal and stemming [68]. Moreover, traditional topic models may struggle with capturing nuanced language patterns, so modern approaches increasingly incorporate deep learning techniques to enhance outcomes [11].

The future of topic modeling lies in integrating these advanced methods, particularly those that can harness contextual information and adapt to domain-specific needs. With ongoing research into dynamic and interactive topic models, there is potential for even more sophisticated tools that can analyze evolving text streams, such as news articles and social media feeds, in real-time [27].

Coherence Score

Topic coherence is critical for assessing the quality of topic models, and Normalized Pointwise Mutual Information (NPMI) is a crucial metric used for this evaluation [61]. NPMI measures the strength of associations between words in a topic, with higher scores indicating stronger, more meaningful relationships. This suggests that the topic is coherent and well-defined [3].

Unlike raw Pointwise Mutual Information (PMI), which measures the association between two variables based on their joint and individual probabilities, NPMI normalizes these scores. This normalization accounts for differences in word frequencies, preventing the metric from being biased by either too common or rare words in the corpus [64].

In practice, the average NPMI score for word pairs within a topic measures that topic's overall coherence. This is useful in validating topic models like LDA [64] or BERTopic, particularly in fields where topic interpretability is crucial, such as academic research, content recommendation, or text summarization.

The NPMI formula is expressed as:

$$NPMI(x,y) = \frac{PMI(x,y)}{-\log(P(x,y))}, \tag{2.6}$$

where:

- $PMI(x,y) = \log(\frac{P(x,y)}{P(x)P(y))})$;
- $P(x,y)$ is the probability of both words occurring together;
- $P(x)$ and $P(y)$ are the probabilities of each word occurring independently.

NPMI scores range from -1 to 1:

- A score of 1 indicates a perfect positive association, where one word predicts the other;
- A score of 0 suggests the words are independent, occurring together no more often than by chance;
- A score of -1 implies a perfect negative association, where one word predicts the absence of the other.

Understanding NPMI scores helps effectively evaluate topic models, ensuring that the topics are coherent, meaningful, and valuable for the intended applications.

2.4 Sentiment Analysis

Sentiment analysis is a critical application of NLP that involves determining the emotional tone behind a body of text [53]. This is a vital tool for understanding human sentiments, especially in large volumes of data, making it crucial for businesses, social media platforms, and customer service interactions.

The core objective of sentiment analysis is to identify and categorize opinions expressed in a text to determine whether the writer's attitude toward a particular topic, product, or service is positive, negative, or neutral. Advanced systems can even detect more nuanced emotions, such as anger, joy, or sadness [41].

Most sentiment analysis systems use ML techniques that require large labeled datasets for training. These systems typically employ supervised learning algorithms, such as logistic regression, support vector machines, or neural networks, to learn from data that has been manually annotated with sentiments [45]. Alternatively, lexicon-based approaches use a dictionary of words that have been pre-assigned specific sentiments. The presence and combination of these words determines the sentiment score of a text. This method does not require training data but often struggles with context and sarcasm, leading to less accurate predictions [70].

Hybrid approaches combine machine learning and lexicon-based methods to leverage both strengths. This can help improve accuracy, especially in scenarios where contextual cues are crucial for determining sentiment [45].

Sentiment analysis finds applications across various industries for different purposes. In business intelligence, companies analyze customer reviews, survey responses, and social media conversations using sentiment analysis to gauge customer satisfaction and identify market trends. Political campaigns and analysts employ sentiment analysis to assess public opinion on policies and candidates. The healthcare industry leverages sentiment analysis to monitor patients' progress in mental health treatments by analyzing their speech or written texts. Furthermore, sentiment analysis plays a crucial role in customer service by automatically categorizing customer feedback, enabling businesses to prioritize responses and understand the overall customer sentiment more effectively [45].

Despite its usefulness, sentiment analysis faces several challenges. Texts often contain sarcasm or implicit meanings that can be difficult for algorithms to interpret correctly, presenting a challenge in handling context and sarcasm. Variations in language, slang, and dialects can affect the results of sentiment analysis tools, making it necessary to adapt systems to specific linguistic contexts [89]. Additionally, some texts may carry ambiguous sentiments or a mixture of positive and negative feelings, complicating the analysis due to ambiguity and neutrality.

Sentiment analysis is leaning towards more sophisticated models that can better handle nuances in language and sentiment. Advances in deep learning and transformer-based models like BERT have already shown promising results in enhancing sentiment analysis systems' accuracy and contextual understanding [16]. Ongoing research aims to refine these technologies, potentially integrating multimodal data (combining text with audio and video) to enrich sentiment detection [40].

2.5 Discussion

We analyze the foundational concepts introduced in this Chapter, including ML, NLP, Language Models, Topic Modeling, and Sentiment Analysis, highlighting their roles and interconnections in advancing AI technologies and vital elements in our study.

The synergy between ML and NLP has not only advanced language understanding but also enabled Language Models to perform a wide array of complex tasks, from translation to text generation. Through ML, especially deep learning techniques, Language Models can identify patterns, predict sequences, and learn from vast datasets. NLP techniques, such as tokenization and contextual embedding, provide the linguistic structure and meaning necessary for interpreting human language effectively.

The introduction of transformers marked a significant shift in NLP, allowing models to handle longer text sequences with greater accuracy, a key asset in analyzing extensive customer feedback datasets. Self-attention mechanisms central to the transformer architecture enable LLMs to focus on the most relevant parts of an input sequence. By capturing intricate patterns within customer feedback, advanced models help to uncover deeper insights, enhancing the interpretive power necessary for reliable, large-scale customer-centric analysis.

Topic modeling is crucial in organizing large text corpora, particularly when understanding the broader themes or topics within a document set is essential. Techniques like LDA [32] and BERT-based topic models such as BERTopic [22] provide a robust way to extract meaningful topics, which can then inform downstream NLP tasks like sentiment analysis. BERTopic and its advanced techniques demonstrate how modern topic models leverage deep learning to offer more accurate and contextually rich topic representations.

2.6 Final Remarks

This Chapter introduced and discussed the foundational concepts and techniques underlying our study, from their essential building blocks to advanced techniques. We presented transformers, topic modeling, and sentiment analysis. These interconnected concepts form the backbone of modern AI applications, particularly in the realm of language understanding and generation.

The rapid evolution of LLMs, demonstrated by models like GPT-3 and Meta's Llama, has unlocked new possibilities in NLP, enabling AI to perform tasks that were previously beyond reach. Meanwhile, topic modeling and sentiment analysis continue to refine how we analyze and interpret large amounts of text, providing deeper insights and enhanced language comprehension.

As we transition into the next Chapter, the focus shifts from foundational concepts to exploring essential related research and developments in topic modeling and sentiment analysis. This sets the stage for an in-depth review of how existing techniques have been applied across various domains and how their integration offers a comprehensive approach to extracting insights from large textual datasets. This exploration provides a solid foundation for understanding the methodologies we employ in this dissertation, where these techniques are applied in practical scenarios like customer feedback analysis.

Chapter 3

Related Work

This Chapter provides an overview of essential related research and advancements in topic modeling and sentiment analysis, which are fundamental techniques for extracting insights from large volumes of textual data.

We begin by exploring various topic modeling techniques, such as LDA, NMF, and modern methods like BERTopic and Top2Vec, each with its own strengths and applications. It was also reviewed advancements in topic modeling that improve clarity and efficiency, highlighting the evolution of these methods in enhancing interpretability and efficiency.

Next, we delve into sentiment analysis, examining both traditional lexicon-based methods and the latest machine learning and deep learning approaches. We also discuss the application of sentiment analysis in various domains, from social media and customer feedback to multilingual and financial contexts. Special attention is given to the growing role of transformer-based models like BERT and RoBERTa, which have significantly advanced the accuracy and effectiveness of sentiment analysis.

Finally, we consider the integration of topic modeling and sentiment analysis, showcasing how the combination of these techniques offers a comprehensive approach to understanding public opinion and uncovering hidden patterns in textual data. Throughout this chapter, we highlight relevant studies that apply these methods in real-world contexts, from education and geopolitical conflicts to cybersecurity and finance, illustrating the wide-ranging impact of these analytical tools.

3.1 Topic Modeling Techniques

Several techniques have been suggested for efficient topic modeling of texts, including LDA, NMF, BERTopic, Top2Vec, and semantics-assisted NMF [38].

Karas *et al.* [32] proposed optimizing LDA and using Top2Vec with doc2vec[1] to extract topics from social media data for rare diseases. Hagen [23] introduced a framework to train and validate LDA for the efficient content analysis of e-petition data.

Shi *et al.* [67] implemented SeaNMF, a model that incorporates word-context semantic correlations, and a sparse variant to achieve better model interpretability. Valdez *et*

[1] Both Top2Vec and doc2vec are NLP techniques used for topic modeling.

al.[76] applied latent semantic analysis to analyze transcripts of the 2016 U.S. presidential debates.

Differently, Chen *et al.* [13] compared LDA and NMF on several public short text datasets and presented Knowledge-guided Non-negative Matrix Factorization for Better Short Text Topic Mining (KGNMF), which leverages external knowledge as a semantic regulator, yielding a time-efficient algorithm.

Abuzayed and Khalifa [2] contrasted LDA and NMF to BERTopic and found that BERTopic using Pre-Trained Arabic Language Models as embeddings generated better results. Egger and Yu [18] evaluated the effectiveness of four topic modeling techniques (LDA, NMF, Top2Vec, and BERTopic) using Twitter data and concluded that BERTopic and NMF are the most effective algorithms for analyzing Twitter data in a social science context.

In the COVID context, Amara *et al.* [5] used Facebook posts in 7 different languages for conducting a LDA-based topic modeling. Results showed that the extracted topics aligned with the chronological development of pandemic-related information and measures taken in various countries.

3.2 Sentiment Analysis Techniques

Soonh *et al.* [71] presented a lexicon-based approach for sentiment analysis of news articles. On the other hand, Park *et al.* [52] proposed LIWC (lexicon-based method) to investigate hotel revisiting behavior.

Like topic modeling, sentiment analysis can be performed by supervised learning algorithms, where words, phrases, and sentences are judged to have a positive, negative, or neutral sentiment. Alternatively, unsupervised learning can be used to recognize patterns and cluster words associated with a specific sentiment.

Rane and Kumar [58] performed a multi-class sentiment analysis on customer feedback tweets for six major US airlines using Doc2vec and ML classification techniques. A case study on customer reviews related to a South African retail bank were conducted by Kazmaier and van Vuuren [33] and showed that custom ML models outperformed pre-trained and commercial tools that are lexicon-based.

Attia *et al.*[7] proposed a language-independent neural network model for sentiment analysis with five layers that do not rely on language-specific features or pre-trained embeddings. The model uses oversampling to address class imbalance and was evaluated on English, German, and Arabic datasets, showing comparable or better results than state-of-the-art methods.

Fuadvy and Ibrahim [20] utilized a multilingual sentiment classifier that collects disaster data from social media in Malaysia, a multicultural and multilingual country, and analyzed the sentiments of people affected by disasters using a deep learning algorithm. A methodology based on BERT was explored by Mehri *et al.* [46] to detect subway-related incidents in tweets. The pre-trained model is context-dependent, fine-tuned for the application, allowing for multilingual and cross-lingual tweet modeling and performs competitively with traditional ML models.

Three studies examined automated sentiment analysis tools in game research. One found that Natural Language Toolkit (NLTK) was the best classifier for game reviews, and identified four causes for incorrect classifications [79]. Another extended the Semantic Orientation Calculator (SO-CAL) sentiment extractor to analyze instant messages in StarCraft 2 [73]. The third used lexical analysis to find most tweets about PUBG were negative, with addiction as a common theme [50].

Sentiment analysis has been widely applied across various domains, leveraging traditional and transformer-based techniques. This selection of studies demonstrates its versatility, addressing challenges in social media sentiment, education, finance, public opinion, and multilingual contexts.

Several studies focus on transformer-based models like BERT and RoBERTa for sentiment analysis. Nuha and Lin [59] utilized BERT and Multilayer Perceptron (MLP) to classify sentiment in user reviews of Taiwan's Social Distancing app, showing that their semi-supervised annotation approach outperformed traditional models. Manik *et al.* [44] also explored the synergy between lexicon-based approaches like SentiStrength and transformer-based models, concluding that incorporating sentiment lexicons improved the accuracy of BERT in sentiment analysis tasks. Similarly, Prasanthi *et al.* [55] applied BERT and RoBERTa to analyze sentiment on social media platforms, emphasizing these models' ability to handle informal language effectively. Vemulapalli and Peddi [78] compared Twitfeel, a traditional sentiment analysis tool, with transformer-based techniques like BERT, concluding that transformers offer clear advantages in accuracy and efficiency.

In the multilingual sentiment analysis space, Shabbir and Majid [66] presented a framework for analyzing Urdu-language data using deep learning techniques like 1D-Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Multilingual-MiniLanguage Model (LM) transformers, achieving significant accuracy improvements. Tian *et al.* [74] introduced An-chiBERT, a pre-trained language model for ancient Chinese tasks such as translation and poem generation outperforming non pre-trained models and BERT. Rath *et al.* [59] extended the application of sentiment analysis to multilingual YouTube live streams, using machine translation and transformers to classify sentiments expressed in multiple languages.

When it comes to financial sentiment analysis, Alissa and Alzoubi [4] fine-tuned BERT and RoBERTa models using a financial sentiment dataset, applying ensemble learning with majority voting. Their results demonstrated enhancements over individual transformer models, showing the value of ensemble approaches in financial sentiment analysis.

Aspect-Based Sentiment Analysis (ABSA) is another significant area, where researchers investigate sentiments related to specific aspects within a text. Yu *et al.* [91] tackled ABSA by proposing a Hierarchical Interactive Multimodal Transformer (HIMT) model to analyze text-image interactions, improving the outcomes on aspect-based sentiment classification tasks. Similarly, Jing and Yang [29] focused on Chinese text sentiment analysis using a Light-Transformer model, optimizing for both efficiency and accuracy in ABSA.

In network public opinion analysis, Dong *et al.* [17] employed BERT to improve sentiment analysis efficiency on social media, while Kumar and Mohan [36] analyzed Twitter data related to the Russo-Ukrainian war using RoBERTa, finding that the majority of public sentiment was negative toward the conflict. Ikeagami *et al.* [25] examined emotional expression in Twitter posts and proposed creating a high-quality Japanese Tweet Emotion corpus using active learning and a fine-tuned Transformer model, this approach improved both emotion analysis and corpus quality.

Together, these studies illustrate the versatility and power of sentiment analysis techniques, particularly the impact of transformer-based models like BERT and RoBERTa. These models demonstrate clear advantages over traditional approaches, whether applied in social media sentiment, multilingual contexts, financial analysis, or specific tasks like ABSA. From improving sentiment classification in user reviews and social media platforms to addressing challenges in multilingual and financial data, transformers continue to push the boundaries of sentiment analysis. These advancements underscore the growing significance of transformer-based techniques in understanding and analyzing public opinion across diverse domains.

3.3 Integration of Topic Modeling and Sentiment Analysis

Sentiment analysis and topic modeling have become powerful elements for understanding public opinion and uncovering hidden concerns across various domains. Several recent studies have applied these techniques to social media data, revealing insights into education, geopolitical conflicts, cybersecurity, and more.

Li *et al.* [39] and Rääf *et al.* [63] both explored the impact of online education, using sentiment analysis and topic modeling to understand the evolving concerns of educators and learners. Li *et al.* [39] focused on sentiments regarding online education during the COVID-19 pandemic, revealing differences between educators and learners on platforms like Reddit. Their study highlighted challenges such as cheating and dissatisfaction with educational strategies, while topic modeling uncovered motivations for using social media. Similarly, Rääf *et al.* [63] analyzed student reviews of MOOCs (Massive Open Online Courses) on Coursera, identifying topics like content delivery and teaching style that shaped learners' satisfaction. Both studies provide valuable insights into the evolving opinions on online education and the factors influencing the online learning experience.

Melo and Figueiredo [15] utilized topic modeling, sentiment analysis, and entity recognition to analyze news articles and tweets in Brazilian Portuguese about the COVID-19 pandemic in Brazil. They found similarities and differences between news and social media, which can aid authorities in decision-making by offering a comprehensive view of public sentiment during the health crisis.

Aslan [6] and Bhardwaj *et al.* [9] examined public sentiment during crises, focusing on geopolitical conflicts and the COVID-19 pandemic, respectively. Aslan [6] used deep learning techniques like the Multistage Feature Extraction using Convolutional Neural Network and Bidirectional Long Short-Term Memory (MF-CNN-BiLSTM) model to analyze tweets about the Ukraine-Russia conflict, capturing public reactions and emotional tendencies. Topic modeling revealed key areas of discussion during the crisis. Bhardwaj *et al.* [9] focused on Twitter data from India during the pandemic, finding that sentiment remained largely positive despite the challenges. Their topic modeling revealed shifts in public discourse across three pandemic waves, from informative messages to broader concerns about tourism and healthcare. Both studies demonstrate the utility of sentiment analysis and topic modeling in understanding public reactions during global crises.

Jang *et al.* [26] explored people's reactions and concerns about COVID-19 in North America, especially in Canada, using LDA for topic modeling and ABSA of related tweets, with negative sentiments related to the overall outbreak, misinformation, and Asians. Positive sentiments were related to physical distancing.

In the realm of artificial intelligence and cybersecurity, Okey *et al.* [51] analyzed tweets about ChatGPT's role in cybersecurity using Valence Aware Dictionary and Sentiment Reasoner (VADER) and RoBERTa, revealing both positive and negative sentiments, especially concerning ChatGPT's potential misuse for hacking. Their use of LDA highlighted public anxiety about cybersecurity threats. Similarly, Nagaraj *et al.* [49] applied sentiment analysis and topic modeling to LinkedIn posts to explore factors behind influential content in professional networks, shedding light on how media content shapes business influence. VADER was also utilized by Dahal *et al.* [14] to study public opinion on climate change through geotagged tweets, employing LDA to reveal negative sentiment with varied topics and country-specific differences. Together, these studies emphasize the growing impact of AI and social media on cybersecurity and professional interactions.

Seki *et al.* [65] also contributed to the discourse by analyzing central bank press conferences using topic-sentiment analysis. Their study of the Bank of Japan's governors showed how sentiment during these conferences influenced economic indicators. By comparing the communication styles of two governors, the study highlighted the significance of tone in shaping economic outcomes.

Cai *et al.* [12] analyzed Reddit's views on ChatGPT and mental health using NLP techniques like BERT (bert-base-multilingual-uncased-sentiment) and BERTopic. They found mostly negative views on AI advice, with some recognizing benefits like affordability and time-saving. Concerns included bad advice, AI replacing therapists, and privacy issues. This study on mental health chatbots showed public worries and hints at AI's promise for better decision-making in various areas.

Together, these studies demonstrate the wide applicability of sentiment analysis and topic modeling in understanding public opinion across education, political conflicts, cybersecurity, professional networking, and economic communications. These techniques not only reveal the emotions behind social media posts but also uncover the topics driving public discourse. From tracking concerns about online education to exploring the role of AI in mental health and cybersecurity, these tools offer valuable insights for policymakers, businesses, and educators. As sentiment and topics shift in response to global events, sentiment analysis and topic modeling provide a critical means of understanding and responding to the evolving public landscape.

3.4 Discussion

The integration of topic modeling and sentiment analysis offers a robust framework for analyzing unstructured textual data, revealing both the thematic structure and the emotional tone embedded within the content. As illustrated by the numerous studies reviewed in this chapter, traditional methods like LDA and NMF continue to provide a solid foundation for topic extraction, particularly when analyzing large corpora with well-defined topics. However, the limitations of these techniques in handling short texts or capturing nuanced semantic relationships have paved the way for the adoption of more sophisticated models such as BERTopic and Top2Vec. These modern approaches harness pre-trained transformer models, allowing for a more contextually rich representation of topics, which has been particularly beneficial in domains such as social media analysis and multilingual datasets.

Similarly, sentiment analysis has evolved from lexicon-based approaches to machine learning and deep learning models, with the advent of transformer-based techniques like BERT and RoBERTa marking a significant shift in the field. These models are better equipped to handle linguistic variability, informal language, and domain-specific sentiment, which is critical in real-time applications such as monitoring public opinion during crises or analyzing consumer feedback in dynamic industries like finance and e-commerce.

The reviewed studies further demonstrate the complementary nature of topic modeling and sentiment analysis. When combined, these methods provide not only a detailed understanding of the topics discussed in a dataset but also the emotional valence associated with these topics. This dual-layered analysis is invaluable in sectors such as education, where understanding both the content and the sentiment of student feedback can shape educational policies, or in geopolitical studies, where gauging public sentiment toward ongoing conflicts can inform diplomatic strategies.

Transformer models, in particular, have shown great promise in enhancing the exactness and depth of sentiment analysis. Their ability to process vast amounts of text and adapt to multilingual contexts allows for a more holistic understanding of public opinion, whether in niche domains like mental health chatbots or broader fields like financial markets. Furthermore, the application of transformers in ABSA illustrates their capability to dissect text at a granular level, revealing both the broader sentiment and the specific aspects contributing to that sentiment.

This thesis contributes to the literature by integrating topic modeling and sentiment analysis to enhance the interpretation of customer feedback, particularly in the context of top mobile apps. It advances the field by demonstrating how the combination of these two techniques provides deeper insights into user sentiment, as opposed to analyzing them in isolation. Additionally, the thesis evaluates the results of fine-tuned models like BERTopic, BERT, and T5 compared to larger pre-trained transformer models such as Llama 3 and Mixtral, offering valuable comparisons for optimizing NLP models in real-world applications. By investigating the impact of topic granularity and sentiment classification, the research adds depth to the understanding of how different configurations improve sentiment analysis outcomes. Furthermore, its application to actual customer reviews bridges the gap between theoretical advancements and practical business intelligence, showing how topic-informed sentiment analysis can aid strategic decision-making across cross-platform datasets.

3.5 Final Remarks

This Chapter reviewed the recent studies and advancements in topic modeling and sentiment analysis, emphasizing their complementary nature in extracting thematic structure and emotional tone from unstructured text data. From traditional methods like LDA and NMF to more advanced techniques such as BERTopic and transformer-based models like BERT and RoBERTa, the evolution of these approaches reflects the growing sophistication of NLP tools in handling complex textual data across various domains. The next Chapter 4 presents our designed methodology for analyzing customer reviews from Amazon, Netflix, and Spotify mobile apps.

Chapter 4

Analyzing Customer Reviews in Textual Data

This Chapter presents our methods for analyzing customer reviews in textual data.

4.1 Overall Methodology

We outline a systematic approach to analyzing customer reviews from three notable mobile apps. More specifically, this study investigates customer reviews of Amazon, Netflix, and Spotify mobile apps from the Google Play Store[1], using a combination of topic modeling and sentiment analysis techniques. By applying advanced natural language processing techniques, our study aims to uncover the main topics discussed in customer reviews and assess the overall sentiment expressed by users across Amazon, Netflix, and Spotify apps. This comprehensive analysis aims to provide valuable insights into customer experiences and opinions on these significant digital platforms.

Figure 4.1 presents the critical steps in our methodology. The research involves **collecting the data** (cf. Section 4.2), **preprocessing** – retrieving the review corpus from the extracted raw data – it(cf. Section 4.3), and addressing **text analysis** (cf. Section 4.4) involving topic modeling and sentiment analysis detection. Our research employs two distinct approaches for conducting topic modeling and sentiment analysis.

The first approach, labeled "Small Fine-Tuned NLP Models" employs BERTopic for topic modeling. We investigate how BERT and T5 models are utilized for sentiment analysis within this approach. Chapter 5 presents our experiments' specific procedures and results to assess this approach.

The second approach, designated "Pre-Trained Large Language Models" employs two large[2] language models, Meta Llama 3 8B and Mixtral 8x7b, for both topic modeling and sentiment analysis. Chapter 6 presents our experiments' specific procedures and results to assess this approach.

[1] Google Play is Google's official app store for Android devices, offering a wide range of apps, games, music, movies, and books. Launched in 2012, it combined several previous services and has since grown to billions of app downloads.

[2] Very large in comparison with BERT and T5, demanding too much computation resources.

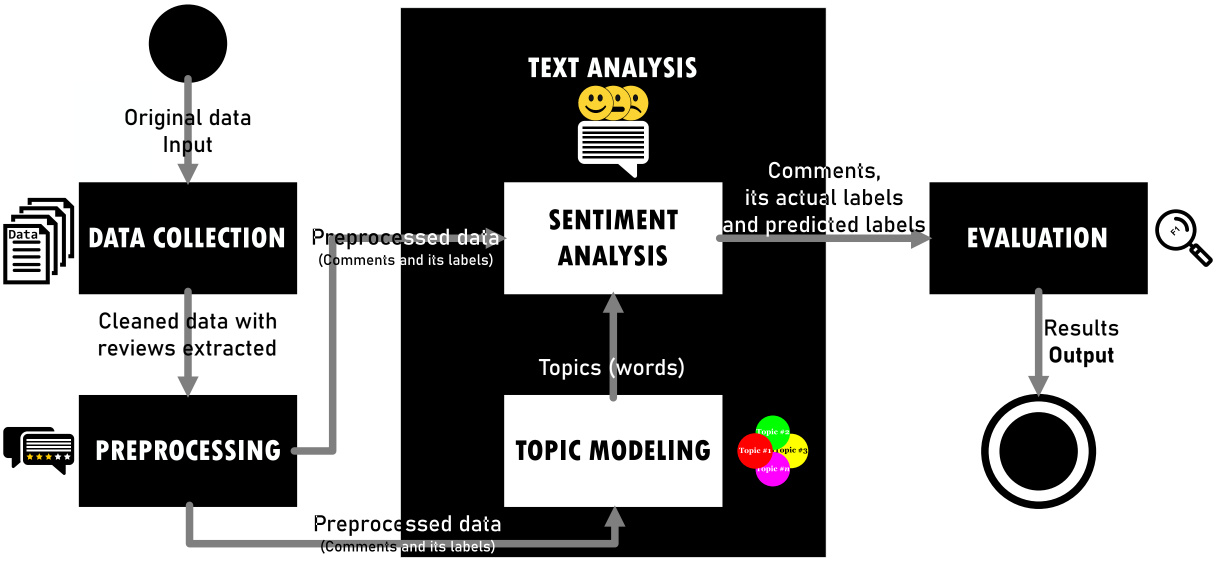


Figure 4.1: Study methodology. This encompasses the processes of data collection, preprocessing, text analysis involving topic modeling and sentiment analysis, and evaluation of results.

Our methodology promotes a comparison between small NLP models like BERTopic and advanced large models such as Meta Llama 3 (8B) and Mixtral (8x7B), providing a comprehensive analysis of topics and sentiments in customer reviews.

The final step of the methodology involves evaluating the effectiveness of all models used (cf. Section 4.5). Overall, we aim to understand and compare the results from the Small Fine-Tuned NLP Models approach with those from the Pre-Trained Large Language Models.

4.2 Obtaining Data

The Amazon App[3] and Netflix App[4] datasets used in this research were sourced from the U.S. Google Play Store through the `google-play-scraper` Application Programming Interface (API)[5]. The Spotify dataset[6], however, was sourced from Kaggle[7] because the API malfunctioned after acquiring the other two datasets.

Several factors justify our research choice of the U.S. Google Play Store. This market's size and diversity make the findings relevant to several app users and potentially reflective of global trends. The substantial volume of user comments provides a rich dataset for comprehensive analysis, enhancing the reliability of results.

Also, models like Meta Llama 3 8B, Mixtral 8x7B, BERT (bert-base-uncased), BERTopic, and T5 (small) are generally optimized for or perform best in English. Some have mul-

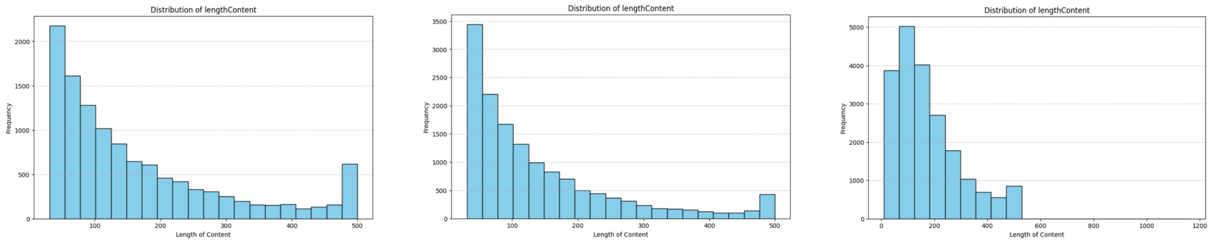
[3] The comments extend from May 5th, 2022 to January 9th, 2024.

[4] The comments range from December 12nd, 2022 to January 26th, 2024.

[5] google-play-scraper: *Google-Play-Scraper provides an API to crawl the Google Play Store for Python. PyPI*, `https://pypi.org/project/google-play-scraper/`. Accessed on June 16, 2024.

[6] The comments span from January 1st, 2022 to July 9th, 2022.

[7] *Spotify App Reviews.* **Kaggle**. Available at: `https://www.kaggle.com/datasets/mfaaris/spotify-app-reviews-2022`. Accessed on March 28, 2024.



(a) Amazon App Dataset (b) Netflix App Dataset (c) Spotify App Dataset

Figure 4.2: Distribution of comments based on their length for each dataset.

tilingual capabilities [16, 57, 75]. By using data from the U.S. Google Play Store, the research leverages these models' strengths, potentially leading to more accurate and insightful results.

These datasets provide a rich source of customer insights and feedback that can inform various aspects of marketing strategy, product development, and customer experience optimization, making them precious resources in a marketing context.

Initially, the original datasets from the API consisted — each (Amazon App and Netflix App) — of 25,000 records, evenly distributed across five sentiment classes: very negative, negative, neutral, positive, and very positive, each class with 5,000 records. These classes were numerically represented from one to five. In contrast, the Spotify dataset contains 61,594 records, with an uneven distribution across the five sentiment labels. The dataset includes 17,653 records for label 1 (very negative), 7,118 records for label 2 (negative), 6,886 records for label 3 (neutral), 7,842 records for label 4 (positive), and 22,095 records for label 5 (very positive). These labels reflect the user's sentiment or opinion about the application, derived from explicit star ratings provided by users, and can be used as a basis for sentiment analysis tasks.

Extraction of Review Data

Privacy considerations were paramount throughout the data extraction process. Measures were taken to anonymize user information and comply with data protection regulations, in the Brazilian scenario the General Personal Data Protection Law[8], ensuring that the analysis remained focused on aggregated trends rather than individual user profiles. In this step, all other useless and non-relevant information related to the scope of this work was discarded.

The extraction phase laid a critical foundation for the subsequent preprocessing and analysis steps, ensuring a rich, reliable, and ethically sourced dataset that reflects the broad spectrum of user experiences across the selected digital platforms.

4.3 Preprocessing Data

Figure 4.2 presents a distribution of comments based on their length for each dataset. Upon examining the distribution of comments by size, we observe that many reviews are empty, containing only the grade. After obtaining the datasets, a crucial step involves preprocessing and cleaning the data. This process removes reviews with empty or excessively short comments, specifically those under 30 characters. Setting a minimum length of 30 characters helps filter out reviews too brief to deliver meaningful feedback or context [19]. By excluding these sparse entries, the dataset becomes more robust, ensuring the analysis is based on more informative and relevant content, enhancing the results' quality and reliability.

Additionally, non-English comments are susceptible to deletion, as it is expected to find comments in languages like Spanish and Hindi on the U.S. Google Play Store. Consequently, the resulting dataset becomes unbalanced across the rating classes.

Because sentiment analysis is performed for three (negative, neutral, and positive) and five (very negative, negative, neutral, positive, and very positive) label scenarios, it is necessary to generate balanced datasets for these two cases. We employed data augmentation techniques using the `nlpaug` [43] library to achieve this balance for the Amazon and Netflix App datasets. This library generates variations of existing text data for training purposes. It provides a range of augmentation techniques such as synonym replacement, random insertion, random swap, and more advanced methods like contextual word embeddings for generating augmented text. By oversampling the minority classes through data augmentation, balanced datasets are created for three and five-label sentiment analysis scenarios[9].

The augmented and balanced data helps mitigate bias and improves the effectiveness of sentiment analysis models trained on these datasets [80].

No data augmentation was required for the Spotify dataset due to its larger size and relatively balanced class distribution. Instead, random undersampling was performed in this case to maintain balance[10].

4.4 Text Analysis

At this stage, the aim is to extract meaningful insights from large volumes of text by leveraging state-of-the-art NLP models and techniques. Two primary tasks are undertaken: **topic modeling** to identify the main themes within the comments and **sentiment analysis** to assess the emotional tone of the users. In the topic-informed sentiment analysis, the predicted topics are used as a complement to the original text, and not as a substitute. We investigate two distinct approaches to address these tasks.

[8] In Portuguese: *Lei Geral de Proteção de Dados Pessoais (LGPD) - Lei 13709/2018*. `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`. Accessed on July 29, 2024

[9] For the Amazon App dataset are generated $24,072$ records for label three and $21,525$ records for label five. In the Netflix case, there are $20,562$ records for label three and $17,720$ for five-label.

[10] There are 20247 for label 3 and 33755 records for label 5.

Small Fine-Tuned NLP Models and Techniques

This approach explored advanced NLP models that were fine-tuned to perform both topic modeling and sentiment analysis on customer reviews. The goal is to derive meaningful insights by utilizing small specialized, task-adapted models.

We employed the BERTopic [22] for topic modeling. While BERT, T5, Meta Llama 3, and Mixtral are examples of pre-trained language models designed to understand and generate text, it is essential to distinguish them from BERTopic [22]. BERTopic is a topic modeling technique that uses embeddings from Transformer-based models such as BERT, but it is not a language model. Therefore, referring to BERTopic as a Transformer-based language model would be inaccurate. Instead, it falls under small fine-tuned NLP techniques that apply these embeddings for specific tasks.

In sentiment analysis, pre-trained transformer models such as BERT Base (uncased) and T5 Small were fine-tuned for sentiment classification. The fine-tuning process adapts these models to specific datasets of customer feedback, allowing them to predict sentiment categories better. The analysis was conducted with and without topic modeling, comparing standalone sentiment predictions to those enhanced by topic insights. Chapter 5 presents the detailed procedures conducted in this evaluation.

Pre-Trained Large Language Models

This approach leverages large pre-trained large language models for topic modeling and sentiment analysis on customer reviews without task-specific fine-tuning.

Using GroqCloud's API with Meta Llama 3 8B and Mixtral 8x7B models, the topic modeling pipeline involved generating topics from user comments, clustering them via *K-means* [84] algorithm, and refining clusters for a concise set of representative topics. Prompts guide the language models throughout this process.

Meta Llama 3 8B is a language model developed by Meta (formerly Facebook) with 8 billion parameters that focuses on efficiency and capabilities in generating human-like text. While Mixtral (8x7B), developed by Mistral AI, is a mixed-architecture model that integrates multiple smaller models (7 billion parameters each) to enhance outcomes and efficiency.

Using three and five sentiment labels, our study compares scenarios with and without topic consideration for sentiment analysis. We employed zero-shot, one-shot, and few-shot learning techniques [31] to enhance model predictions. Chapter 6 presents an in-depth explanation of specific procedures conducted in this evaluation.

4.5 Evaluation Methods and Metrics

The evaluation of this study focused on assessing the effectiveness of both Small Fine-Tuned NLP Models and Pre-Trained Large Language Models in analyzing customer reviews from Amazon, Netflix, and Spotify.

Key metrics such as Accuracy, Precision, Recall, and F1-score [8] were calculated to evaluate the models. The F1-score is given particular attention due to its ability to balance Precision and Recall, providing a comprehensive understanding of each model's capability to predict sentiment across different classes.

For topic modeling, the *coherence score* [34] was used to assess how well the topics generated by the models make sense semantically and how they reflect the main themes within the dataset. The evaluation was conducted in two critical scenarios: with and without the integration of topic modeling into the sentiment analysis process.

Evaluating sentiment analysis with and without topic modeling integration helps us understand the added value of contextual topic information in sentiment predictions.

In the first scenario, the models were evaluated purely on their ability to classify sentiment based on textual data without considering the topics identified. This served as a baseline, enabling a straightforward comparison of model effectiveness in assessing the inherent ability to interpret emotional tones in customer feedback based solely on sentiment detection.

In contrast, integrating topic modeling into the sentiment analysis process is hypothesized to enhance the accuracy and relevance of sentiment classifications by grounding them in specific user concerns. This approach identifies the main themes in customer reviews, allowing us to assess whether understanding these topics improves the depth and certainty of sentiment analysis. We analyzed the impact of this integration to determine if models perform better when informed by the underlying topics discussed in the reviews.

Furthermore, the Small Fine-Tuned NLP Models were compared to the Pre-Trained Large Language Models to assess how different model architectures and configurations affect the results. For the fine-tuned models, the evaluation examines how task-specific adaptations, such as fine-tuning on customer review datasets, improve sentiment classification. In contrast, we assessed the pre-trained models' ability to handle large-scale data without specific fine-tuning, using techniques like zero-shot, one-shot, and few-shot learning.

Finally, statistical significance tests were applied – when appropriate – to ensure that any observed differences in results between models and approaches are meaningful. These tests confirm whether the integration of topic modeling or the use of different learning configurations provides substantial improvements in sentiment analysis accuracy rather than being a result of random variation.

Chapter 5 and Chapter 6 present the achieved results.

4.6 Final Remarks

This Chapter outlined our methodology to provide a systematic approach to analyzing customer reviews using state-of-the-art techniques. Evaluating different models, datasets, label scenarios, and configurations, remarkably the comparison between small fine-tuned and pre-trained large models, enhances our understanding of best applying NLP models to sentiment classification tasks. Our approach establishes a foundation for detailed analysis, using advanced techniques to enable a deeper evaluation of the most suitable settings by conducting this analysis across different configurations.

The next Chapter, Small Fine-Tuned NLP Models and Techniques (Chapter 5), describes the research approach that uses advanced NLP models to analyze customer reviews and compares the sentiment predictions obtained without topics to those made with the help of topics. This comparison highlights how the inclusion of topics can enhance sentiment prediction metrics. These findings complement the investigation presented in Chapter 6 concerning Pre-Trained Large Language Models that offer insights into the strengths and limitations of leveraging powerful, general-purpose models without additional fine-tuning. We evaluate whether specialized training provides advantages over general-purpose pre-trained models.

Chapter 5

Evaluating Small Fine-Tuned NLP Models and Techniques

The primary objective of this chapter is to evaluate small fine-tuned NLP models for sentiment analysis and topic modeling, assessing how well they perform in identifying nuanced patterns within customer feedback. By exploring models like BERT and T5, alongside different configurations for topic modeling with BERTopic, this chapter aims to determine the efficacy of fine-tuned methods across various classification and topic modeling settings.

The chapter is organized into the following sections: Section 5.1 (Specific Procedures for Topic Modeling and Sentiment Analysis Tasks) describes the topic modeling and sentiment analysis methods applied, including model architecture, training settings, and evaluation metrics.

In Section 5.2, the outcomes of sentiment analysis and topic modeling are presented, highlighting comparative results across different labeling schemes and modeling approaches.

Then the Discussion (cf. Section 5.3), analyzes the comparative findings, emphasizing the impact of each method on model results and coherence, and draws implications for optimal classification and topic modeling strategies.

The chapter concludes with the Final Remarks (5.4), with a summary of key insights and their significance for transitioning to pre-trained large language models in the following chapter.

5.1 Specific Procedures for Topic Modeling and Sentiment Analysis Tasks

In Figure 5.1 a visual representation of the different settings explored is presented. This will be further explained in the following subsections inside this one.

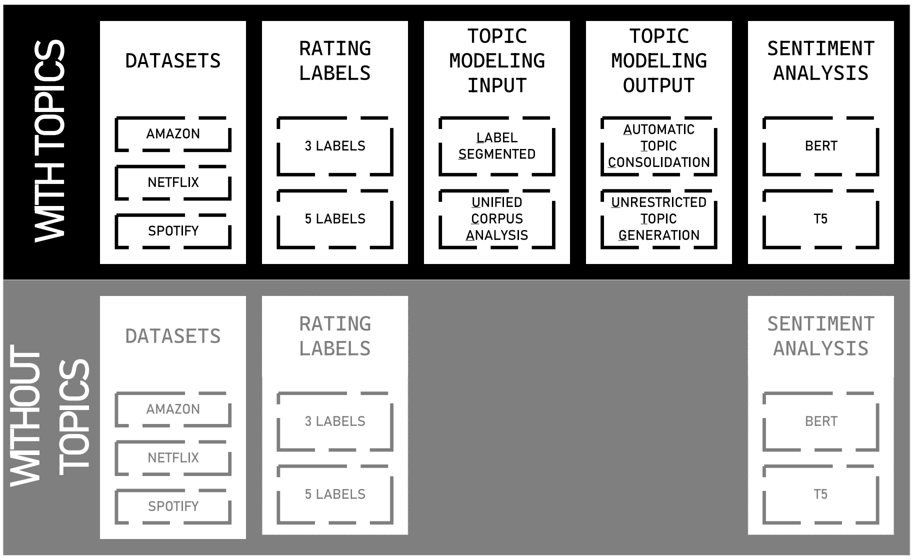


Figure 5.1: Different Configurations For The Small Fine-Tuned NLP Models With Topics and Without Topics.

5.1.1 Topic Modeling

This step — in this specific scenario — was executed using BERTopic [22], an algorithm that clusters similar topics in large text corpora using contextualized embeddings[1]. BERTopic was chosen for its natural language processing capabilities, dynamic topic modeling, and suitability for the diverse and complex nature of customer reviews, enabling more nuanced and accurate analysis of customer feedback [22].

In BERTopic, it is possible to determine the number of top words that should represent each topic. In this context, we set this value for 25; this specific number allows for capturing a wide range of the most relevant and descriptive terms associated with that topic while facilitating clear topic interpretability[2]. Values further than that return incoherent topics[3].

The topic modeling process with BERTopic involves determining the optimal number of topics to capture meaningful themes in the data. In this research, two approaches are employed. The first approach, called Automatic Topic Consolidation (ATC), begins with a smaller initial number of topics, which is then refined and automatically reduced using HDBSCAN [22], a clustering algorithm that consolidates similar topics. This consolidation typically results in a more focused topic set, often producing between 10 and $1,225$ topics. Alternatively, the Unrestricted Topic Generation (UTG) approach generates a larger number of topics by bypassing any reduction techniques. This results in a higher topic count that can range widely, from 428 to over $3,911$ topics. The UTG approach

[1] Embeddings in NLP represent words or phrases as numerical vectors, allowing for capturing the semantic meaning and relationships between them. They are learned from extensive text data using unsupervised machine learning algorithms and used as input for various NLP tasks to generate more correct results.

[2] Regarding topic modeling, specifically, many decisions are made by human judgment, there is no playbook for conducting topic modeling. In this specific case, it was noticed that choosing small values of top words would suppress relevant information specific to some comments.

[3] In this particular case with the use of BERTopic.

provides a more granular and expansive set of topics, while ATC delivers a more concise set of refined topics.

Topic modeling can be conducted in two main ways, each with distinct effects on the granularity and focus of the topics generated. One approach, known as Label-Segmented (LS), involves dividing the data according to pre-existing labels, such as rating classes (e.g., 1 to 5-star ratings). By segmenting the data in this way, topic modeling is performed within each label group separately. This allows for a more targeted analysis, where specific topics emerge within each sentiment or rating class, offering insights into themes that may vary across different sentiment intensities or customer feedback ratings.

In contrast, Unified Corpus Analysis (UCA) treats the entire dataset as a single corpus, ignoring any individual label distinctions during the topic modeling process. This method enables the identification of overarching topics that span the entire corpus, capturing broad themes that apply universally across all sentiment levels or rating classes. UCA, therefore, provides a more generalized view of the topics within customer feedback, revealing common themes that are not specific to any particular sentiment level.

Overall, LS allows for label-specific topic analysis, potentially exposing unique themes within each sentiment level, while UCA reveals universal topics that characterize the customer feedback as a whole.

After fitting the BERTopic model to the preprocessed text data (the provided dataset), it generates two main outputs. The first output is the topic assignments for each document[4] in the corpus. Essentially, it determines which topics are present in each document, the combination of topics and mapped documents (comments) are provided in the sentiment analysis step. The second output is the probability or confidence level associated with each topic assignment. This indicates how strongly a particular topic is represented within a given document. An example is given as follows:

- Comment: "*furthermore, you didn't realize what i will do without it? fair prices & very large delivery, what exactly can you want?*"
  - Topics: ('*fast delivery want*', '*realize fair prices*', '*realize fair*', '*prices large delivery*', '*prices large*', '*large delivery exactly*', '*large delivery*', '*know fantastic prices*', '*know fantastic*', '*furthermore didn realize*', '*furthermore didn*', '*want amazing*', '*fantastic prices super*', '*fantastic prices*', '*fair prices large*', '*exactly want amazing*', '*want amazing don*', '*delivery want*', '*didn realize*', '*didn realize fair*', '*delivery exactly*', '*amazing don know*', '*amazing don*', '*delivery exactly want*', '*don know fantastic*')
  - Probability / Confidence level: (0.075, 0.072, 0.070, 0.069, 0.067, 0.061, 0.060, 0.058, 0.056, 0.053, 0.052, 0.051, 0.047, 0.047, 0.042, 0.041, 0.040, 0.040, 0.036, 0.035, 0.034, 0.029, 0.028, 0.015, 0.009)

Finally, a "coherence" score is calculated to reflect how healthy words within each topic relate to each other. This score is computed using the Gensim library [60], where the coherence model evaluates the quality of the topics generated by measuring the degree

[4] In this case it refers to the comments, a document refers to a single piece of text that you want to analyze and extract topics from.

of semantic similarity among the top words in each topic. The higher the coherence score, the more meaningful and interpretable the topics are considered to be.

5.1.2 Sentiment Analysis

The current approach (Small Fine-Tuned NLP Models and Techniques) computed sentiment analysis via BERT base (uncased) and T5 Small transformer models from Hugging Face's transformers library [88]. The selection of BERT base and T5 Small models for sentiment analysis in our study was primarily guided by the balance between computational efficiency and model effectiveness [16, 57], given the constraints of our computational resources. More specifically, these models were chosen to avoid excessive computational load, as larger models, such as BERT Large or T5 Base, would have required significantly more memory and processing power, potentially exceeding the capacity of our available hardware.

Sentiment analysis can be performed both with and without specific topics. When analyzing sentiments with topics, we evaluate all configurations based on the output generated by BERTopic with the comments as well. This involves looking at both LS and UCA, along with considering the number of topics, which is determined by the ATC and the UTG.

These pre-trained models were fine-tuned specifically for the sentiment analysis task by training them on labeled data, where each review – and topics (in the case with topics) – had a predefined sentiment label. The fine-tuning process involved adapting models to recognize patterns in text data related to sentiment categories by minimizing the error between predictions and actual labels. This training allowed the models to adjust their parameters and learn to associate specific phrases and expressions with sentiments, enhancing their ability to analyze customer reviews effectively.

The data (from all datasets) was split into 80% for training and 20% for testing. Fine-tuning occurred over five epochs with a batch size of 16, a learning rate of 7e-05, 500 warmup steps, and a weight decay of 0.001, with evaluations at each epoch's end. The hyperparameters were selected based on a combination of best practices in the literature [72] and the constraints of our computational resources while maximizing the effectiveness of our sentiment analysis models. This configuration ensures that our models are both practical for our computational environment[5] and capable of achieving high accuracy in sentiment analysis tasks.

The effectiveness across sentiment classes was evaluated using accuracy, precision, recall, and F1 score metrics. This analysis provides insights into each model's results and identifies areas for improvement, making this a comprehensive approach that combines advanced transformer models, targeted fine-tuning, and detailed evaluation. It enables effective sentiment prediction in customer reviews.

[5] Computational Environment: Intel(R) Xeon(R) Central Processing Unit (CPU) E5-2640 v2 @ 2.00GHz with 16 cores, 16 GB RAM, and NVIDIA RTX A6000 Graphics Processing Unit (GPU) with 49 GiB memory.

5.1.3 Null Hypothesis

To rigorously assess the effects of topic modeling and model selection on sentiment classification results, we apply statistical hypothesis testing throughout this study. Specifically, we use the Wilcoxon Signed-Rank Test, a non-parametric test suitable for paired comparisons in this context, to evaluate whether topic modeling leads to significant differences in F1 scores across different models, approaches, and labeling schemes.

The hypotheses tested include:

- **Three-label classification under UCA**:
  - **Model comparison**: Adding topics does not lead to a significant difference in the F1 scores of BERT and T5 in the 3-label classification setting under UCA.
  - **Topic modeling approach**: Adding topics does not produce a significant difference in F1 scores between the UTG and ATC scenarios in the 3-label classification setting under UCA.
- **Three-label classification under LS**:
  - **Model comparison**: Adding topics doesn't make a significant difference in F1 scores between the BERT and T5 in the 3-label classification setting under LS.
  - **Topic modeling approach**: Adding topics does not produce a significant difference in F1 scores between the UTG and ATC scenarios in the 3-label classification setting under LS.
- **Five-label classification under UCA**:
  - **Model comparison**: Adding topics fails to create a significant difference in the F1 score between BERT and T5 in the 5-label classification setting under UCA.
  - **Topic modeling approach**: Adding topics does not produce a significant difference in F1 scores between the UTG and ATC scenarios in the 5-label classification setting under UCA.
- **Five-label classification under LS**:
  - **Model comparison**: Adding topics doesn't make a significant difference in F1 scores between the BERT and T5 in the 5-label classification setting under LS.
  - **Topic modeling approach**: Adding topics does not produce a significant difference in F1 scores between the UTG and ATC scenarios in the 5-label classification setting under LS.

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Amazon	T5	0.603	0.614	0.603	0.561
	BERT	0.617	0.639	0.617	0.594
Netflix	T5	0.553	0.548	0.553	0.494
	BERT	0.629	0.632	0.629	0.625
Spotify	T5	0.684	0.69	0.684	0.684
	BERT	0.699	0.7	0.699	0.687

Table 5.1: Evaluation Metrics for BERT and T5 Models in 3-Label Sentiment Analysis Without Topics Across Datasets.

These hypotheses allow to statistically examine whether topic modeling and model choice influence sentiment classification effectiveness. Introducing these tests at this stage provides a structured foundation for interpreting results with greater clarity.

5.2 Results

This section presents the results of our application of small fine-tuned NLP models and techniques to enhance sentiment analysis across multiple datasets.

The focus is on the comparative results of transformer-based models —BERT and T5— across various classification scenarios, both with and without the integration of topic modeling. These models serve as foundational elements for understanding the impact of incorporating advanced techniques like topic modeling to improve the granularity and accuracy of sentiment predictions.

We explore distinct topic modeling approaches. Their influence on sentiment classification will be analyzed.

5.2.1 Results of Sentiment Analysis Without Topics

These initial results provide a vital benchmark for assessing the impact of incorporating topic modeling. Given the very small sample size in this study, we prioritized descriptive statistics as our primary analysis method. This approach allows us to summarize and describe key characteristics of the data without applying inferential statistical tests that may be unreliable with limited samples. By focusing on measures such as the mean, we can effectively capture and compare central tendencies and variability in the outcomes of the models across different scenarios with three and five labels. This method ensures a more reliable analysis, reducing the risk of drawing misleading conclusions from the small sample size, while maintaining a consistent overall average.

Table 5.1 reports the sentiment analysis effectiveness of BERT and T5 models on a three-class labeling configuration without topic modeling. Table 5.2 presents the same analysis for a five-class labeling scheme. These baseline result metrics highlight the models' general sentiment analysis capabilities across the different datasets.

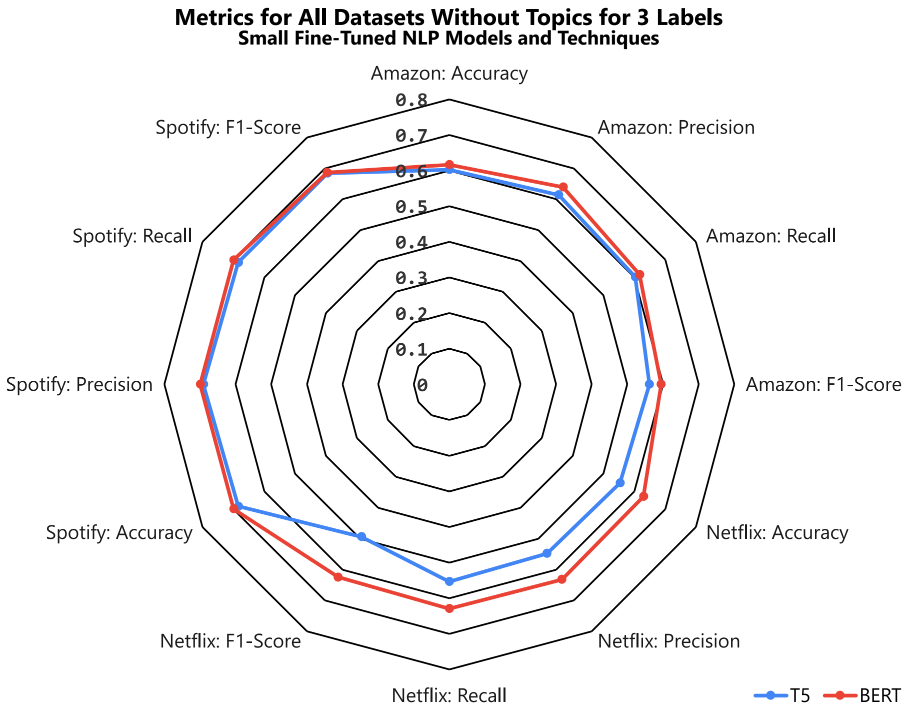


Figure 5.2: Radar Chart With Metrics for BERT and T5 Models Across Datasets for 3-Label Sentiment Analysis Without Topics.

Three-labels results

For the three-class labeling scheme, BERT consistently outperforms T5 across all datasets. The most significant gap between the F1 scores of T5 and BERT is observed in the Netflix dataset, where BERT scores 0.625 compared to T5's 0.494. BERT achieves the highest F1-score of 0.687 on the Spotify dataset, while T5 underperforms with an F1 of 0.561 on the Amazon dataset. These observations are illustrated in Figure 5.2. The overall average F1 score across all models is 0.607. For BERT models specifically, the average F1 score is 0.635, while for T5 models it is 0.58. Notably, the highest F1 score was achieved by BERT.

Five-labels results

In the five-class labeling scenario, detailed in Table 5.2 and Figure 5.3, the gap between T5 and BERT F1-scores is minimal for the Amazon dataset, differing by only 0.004. For the Netflix dataset, BERT slightly outperforms T5 with an F1-score of 0.48 compared to T5's 0.477, both representing the lowest score observed across all datasets and models in this scenario. The highest F1-score, 0.522, was achieved by BERT on the Spotify dataset, while T5 achieved 0.519 on the Amazon dataset. The average F1-score for BERT is 0.506 and for T5 is 0.502, resulting in an overall average F1-score of 0.504.

5.2.2 Results of Topic-Informed Sentiment Analysis

The effect of topic modeling on sentiment analysis was examined using two approaches: UTG and ATC, across Amazon, Netflix, and Spotify datasets.

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Amazon	T5	0.527	0.52	0.527	0.519
	BERT	0.527	0.512	0.527	0.515
Netflix	T5	0.481	0.476	0.481	0.477
	BERT	0.472	0.515	0.472	0.48
Spotify	T5	0.522	0.512	0.522	0.509
	BERT	0.531	0.527	0.531	0.522

Table 5.2: Evaluation Metrics for BERT and T5 Models in 5-Label Sentiment Analysis Without Topics Across Datasets.

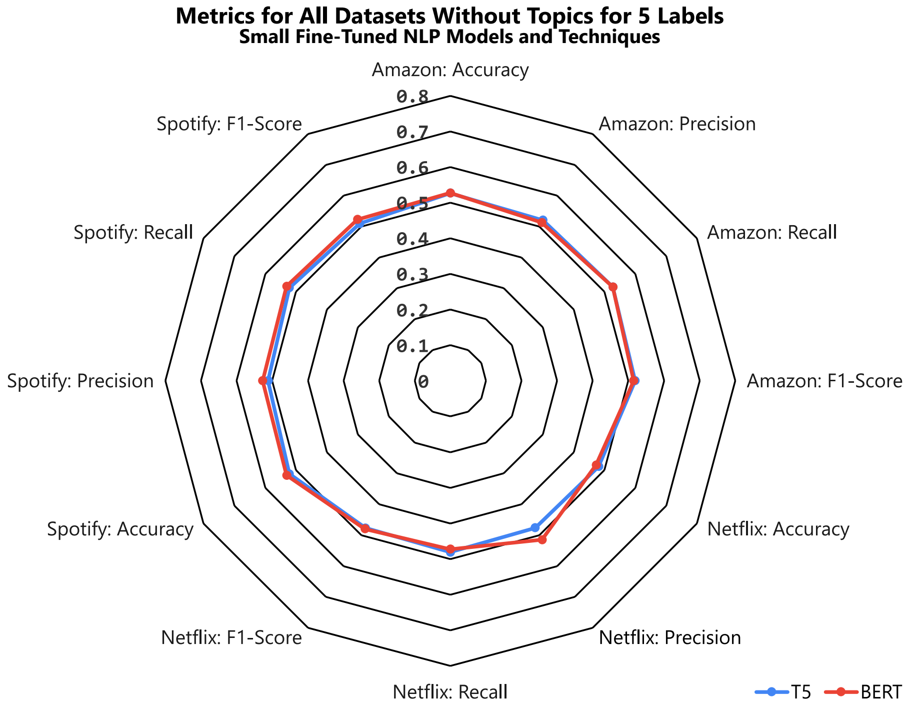


Figure 5.3: Radar Chart With Metrics for BERT and T5 Models Across Datasets for 5-Label Sentiment Analysis Without Topics.

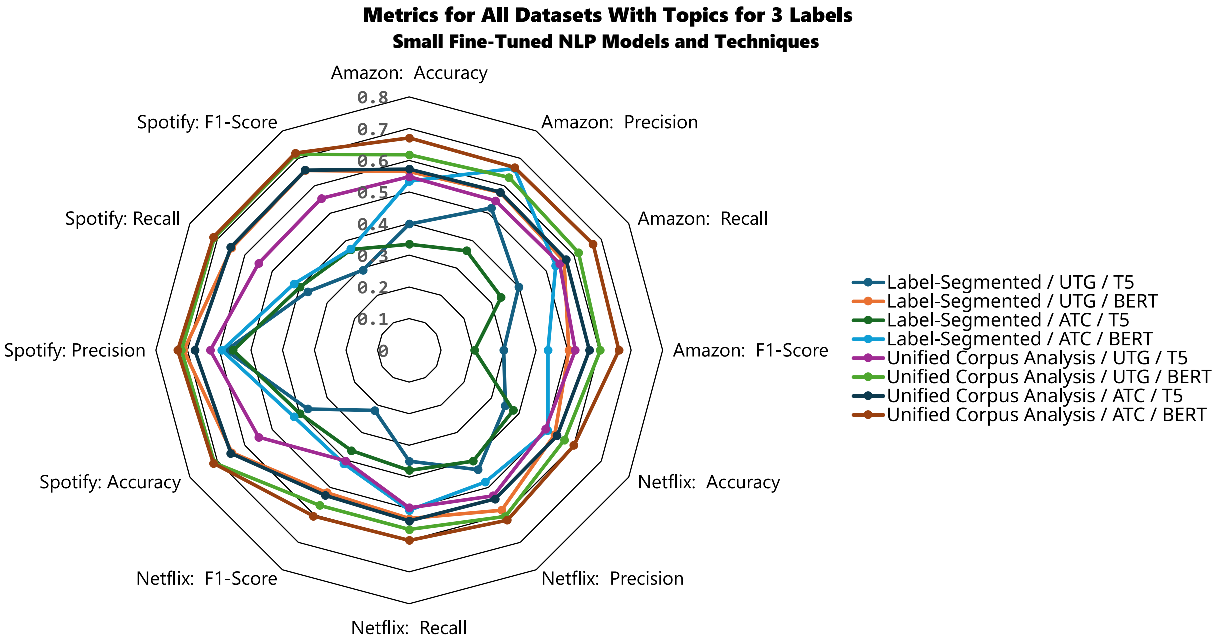


Figure 5.4: Radar Chart With Metrics for BERT and T5 Models Across Datasets for 3-Label Sentiment Analysis With Topics.

Three-labels results

Figure 5.4 illustrates the metrics for sentiment analysis models integrated with UCA and LS topic modeling techniques under the ATC and UTG scenarios using a three-class approach. These results provide a visual comparison of model results across different topic modeling configurations. Complementing this, Table 5.3 and Table 5.5 present the detailed metrics for each sentiment analysis model, allowing for a more granular examination of UCA and LS outcomes across both ATC and UTG setups.

Table 5.3 presents no segmentation (UCA) for the three-class labeling scheme, the Spotify dataset attained the highest F1 score of 0.719 with the BERT model under the ATC scenario, closely followed by 0.714 under UTG. Similarly, the Amazon dataset scored 0.662 with BERT under ATC. Under UTG, the Amazon dataset scored 0.603 with BERT. The Netflix dataset's best F1-score was 0.565 with BERT under UTG. The T5 model resulted in lower overall scores, achieving F1 scores of 0.657, 0.569, and 0.529 for the Spotify, Amazon, and Netflix datasets, respectively, under the ATC scenario. Under UTG, T5 scored 0.554 for Spotify, 0.524 for Amazon, and 0.403 for Netflix. The overall F1 average was 0.592, with BERT achieving the highest average of 0.645, while T5 had an average of 0.539. For the UTG and ATC scenarios, the F1 averages were 0.56 and 0.623, respectively, with ATC outperforming UTG.

Considering the null hypothesis "Adding topics does not lead to a significant difference in the F1 scores of BERT and T5 in the three-label classification setting under UCA", when comparing BERT and T5 in this setting, the Wilcoxon Signed-Rank Test yielded a p-value[6] of 0.043, assuming the null hypothesis is true. Since this p-value is less than the

[6] Representing the probability of observing a test statistic as extreme as, or more extreme than, the

Dataset	Approach	Accuracy	Precision	Recall	F1
Amazon	T5 / UTG	0.548	0.545	0.548	0.524
	BERT / UTG	0.617	0.63	0.617	0.603
	T5 / ATC	0.572	0.576	0.572	0.569
	BERT / ATC	0.67	0.666	0.67	0.662
Netflix	T5 / UTG	0.497	0.53	0.497	0.403
	BERT / UTG	0.566	0.606	0.566	0.565
	T5 / ATC	0.539	0.542	0.539	0.529
	BERT / ATC	0.6	0.619	0.6	0.605
Spotify	T5 / UTG	0.549	0.628	0.549	0.554
	BERT / UTG	0.711	0.72	0.711	0.714
	T5 / ATC	0.651	0.677	0.651	0.657
	BERT / ATC	0.714	0.731	0.714	0.719

Table 5.3: Evaluation of BERT and T5 Models in 3-Label Topic-Informed Sentiment Analysis with UTG and ATC Using UCA Across Datasets.

significance level of 0.05, we reject the null hypothesis, providing statistically significant evidence that there is a difference between the BERT scores and T5 scores.

In contrast, for the comparison between UTG and ATC scenarios, the null hypothesis was that adding topics does not result in a significant difference in F1 scores between these scenarios under the same classification setting. Here, the Wilcoxon Signed-Rank Test produced a p-value of 0.094. We fail to reject the null hypothesis since this p-value exceeds the 0.05 significance level. Thus, results are not significantly different between the UTG and ATC scenarios.

The topic coherence scores (Table 5.4) remained consistent across the sentiment classes, hovering around 0.574 for Amazon, 0.503 for Netflix, and 0.56 for Spotify. This stability highlights the ability of the ATC approach to produce fewer, more focused themes without compromising interpretability. In contrast, the UTG method generated a significantly larger number of topics - $2,884$, $2,517$, and $2,412$ for Amazon, Netflix, and Spotify, respectively. ATC effectively consolidated these themes to 162, 496, and 36 topics while maintaining the coherence levels. This demonstrates ATC's effectiveness in generating a more streamlined set of interpretable topics compared to the more expansive UTG approach. In Figure 5.5 there are the coherence values across different situations and in Figure 5.6 the number of topics generated.

Table 5.5 presents label-segmented topic modeling results for three classes; the best results were observed for the Spotify, Netflix, and Amazon datasets under the UTG scenario with the BERT model, achieving F1 scores of 0.656, 0.519, and 0.504, respectively. Conversely, the T5 model yielded the poorest results, with scores as low as 0.292 and 0.219 for the Spotify and Netflix datasets under UTG, and 0.206 for the Amazon dataset under ATC, the lowest across all scenarios. The average F1 score was higher for BERT at 0.483, compared to 0.291 for T5. Similarly, the UTG scenario had a higher average F1 score of 0.415, compared to 0.36 for ATC. Overall, the average F1 score across models and scenarios was 0.387.

one obtained.

Dataset	Approach	Coherence	Topics
Amazon	UTG	0.574	2,884
	ATC	0.574	162
Netflix	UTG	0.503	2,517
	ATC	0.503	496
Spotify	UTG	0.56	2,412
	ATC	0.56	36

Table 5.4: Topic Coherence and Number of Topics for UCA Analysis in a 3-Label Classification Context.

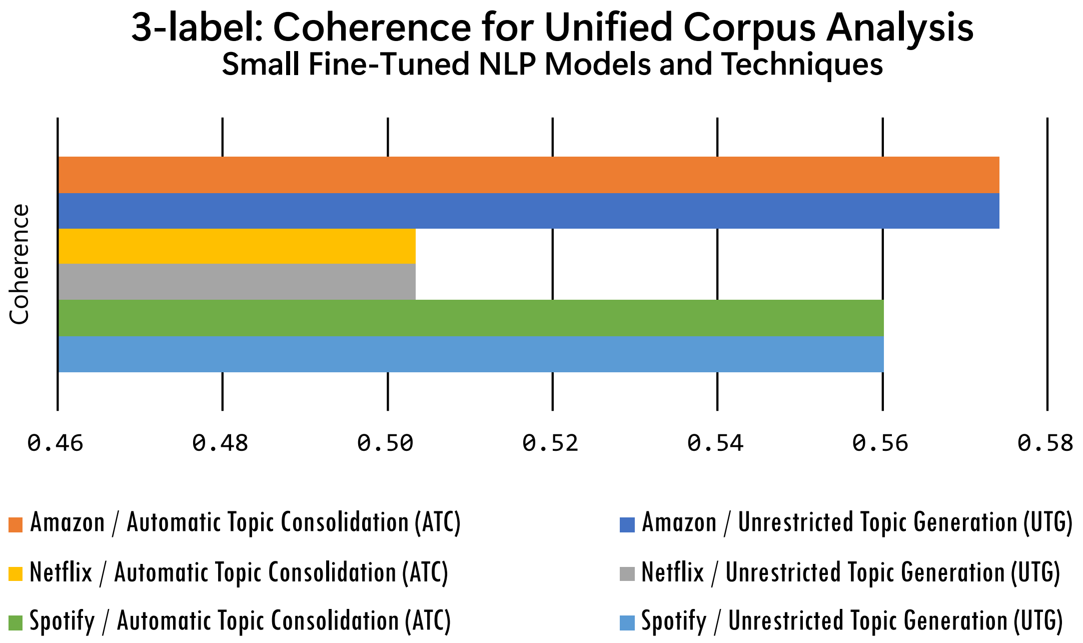


Figure 5.5: Bar Graph of Coherence Scores for All Datasets and Approaches in 3-Label Classification Using UCA.

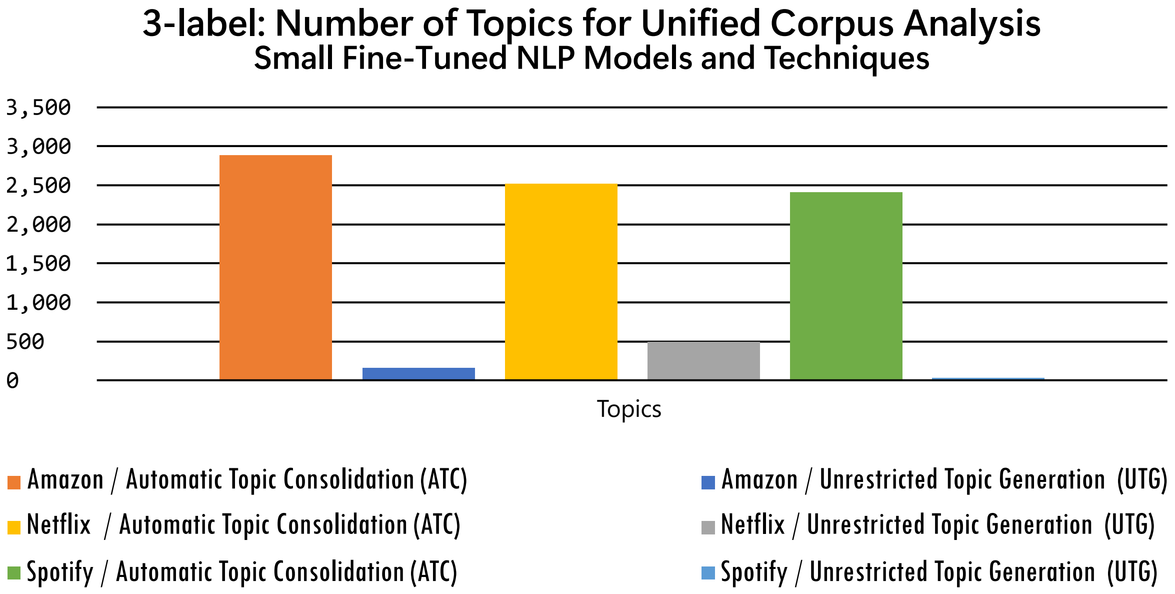


Figure 5.6: Bar Graph of the Number of Topics for All Datasets and Approaches in 3-Label Classification Using UCA.

Dataset	Approach	Accuracy	Precision	Recall	F1
Amazon	T5 / UTG	0.399	0.519	0.399	0.298
	BERT / UTG	0.564	0.575	0.564	0.504
	T5 / ATC	0.335	0.363	0.335	0.206
	BERT / ATC	0.534	0.663	0.534	0.438
Netflix	T5 / UTG	0.35	0.435	0.35	0.219
	BERT / UTG	0.531	0.583	0.531	0.519
	T5 / ATC	0.379	0.404	0.379	0.366
	BERT / ATC	0.505	0.48	0.505	0.413
Spotify	T5 / UTG	0.37	0.579	0.37	0.292
	BERT / UTG	0.647	0.712	0.647	0.656
	T5 / ATC	0.399	0.556	0.399	0.367
	BERT / ATC	0.42	0.592	0.42	0.369

Table 5.5: Evaluation of BERT and T5 Models in 3-Label Topic-Informed Sentiment Analysis with UTG and ATC Using LS Across Datasets.

The null hypothesis stating that “Adding topics does not make a significant difference in F1 scores between the BERT and T5 in the three-label classification setting under LS” is rejected. This conclusion is based on the Wilcoxon Signed-Rank Test, which yielded a p-value of 0.043, indicating a significant difference between the models. Conversely, when comparing the UTG and ATC scenarios, the Wilcoxon Signed-Rank Test produced a p-value of 0.562. This result suggests that the null hypothesis — “Adding topics does not produce a significant difference in F1 scores between the UTG and ATC scenarios in the three-label classification setting under LS” — cannot be rejected, indicating no significant difference between these scenarios.

Topic coherence remained stable across the datasets, with Amazon scoring around 0.65, Netflix achieving 0.729, 0.698, and 0.65 for negative, neutral, and positive sentiments, respectively, and Spotify scoring 0.625, 0.706, and 0.658 for the same sentiments as observed in Table 5.6 and in Figure 5.7. The number of topics varied significantly between the modeling approaches. ATC produced more focused topic distributions: Amazon had 94 topics for negative, 17 for neutral, and 269 for positive sentiments; Netflix had 201, 101, and 340 topics, respectively; and Spotify had 96, 282, and 16 topics. In total, ATC generated 380 topics for Amazon, 642 for Netflix, and 394 for Spotify. In contrast, the UTG approach generated a much larger number of topics, resulting in a broader but more dispersed thematic structure. For Amazon, UTG produced 940 topics for negative, $1,069$ for neutral, and $1,036$ for positive sentiments, totaling $3,045$ topics. Netflix had 872, 935, and 865 topics, totaling $2,672$, and Spotify had 812, 841, and 800 topics, culminating in $2,453$ topics. These results highlight the trade-off between the concentrated focus of ATC and the broader coverage of UTG. In Figures 5.8 and 5.9 is possible to observe the number of topics per class and the total number of topics respectively.

The coherence values between ATC and LS did not change in this 3-label setting. In comparison to UCA, the coherence values for LS are higher. Specifically, the mean coherence values for LS are 0.649 for Amazon, 0.692 for Netflix, and 0.663 for Spotify. These values exceed the corresponding UCA values for the same datasets, which are 0.574,

Dataset	Label	UTG		ATC	
		Coherence	Topics	Coherence	Topics
Amazon	0 (negative)	0.647	940	0.647	94
	1 (neutral)	0.653	1,069	0.653	17
	2 (positive)	0.646	1,036	0.646	269
Netflix	0 (negative)	0.729	872	0.729	201
	1 (neutral)	0.698	935	0.698	101
	2 (positive)	0.65	865	0.65	340
Spotify	0 (negative)	0.625	812	0.625	96
	1 (neutral)	0.706	841	0.706	282
	2 (positive)	0.658	800	0.658	16

Table 5.6: Topic Coherence and Number of Topics for LS Analysis in a 3-Label Classification Context.

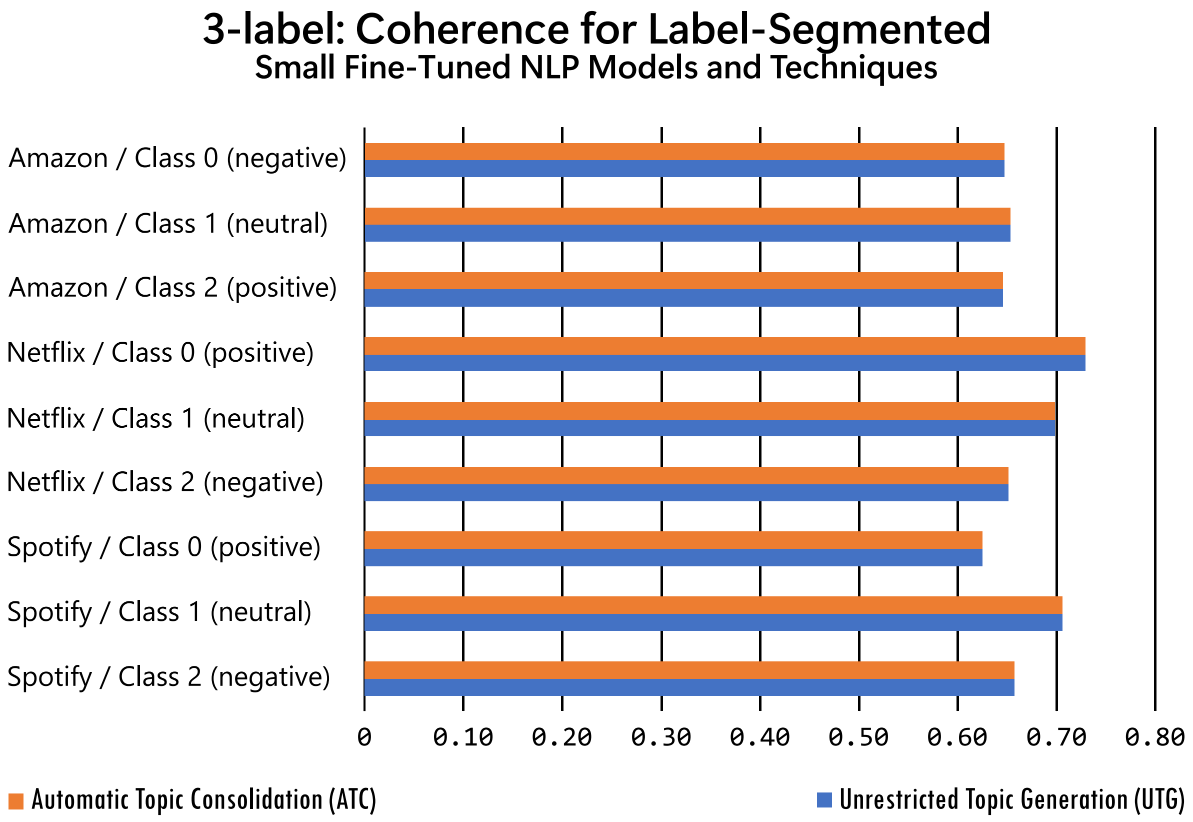


Figure 5.7: Bar Graph of Coherence Scores for All Datasets and Approaches in 3-Label Classification Using LS Topic Modeling.

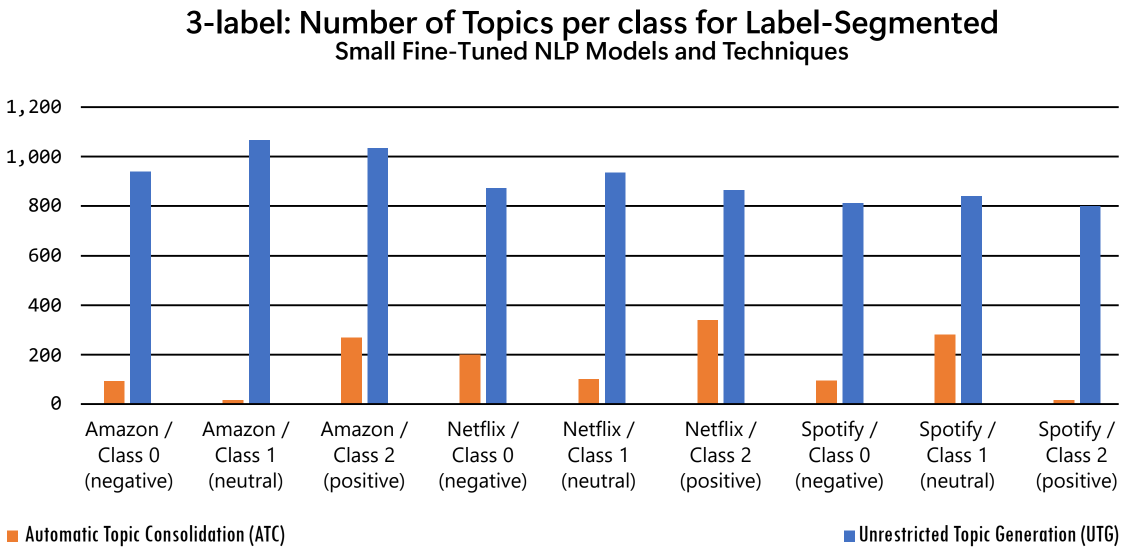


Figure 5.8: Bar Graph of the Number of Topics per Class for All Datasets and Approaches in 3-Label Classification Using LS Topic Modeling.

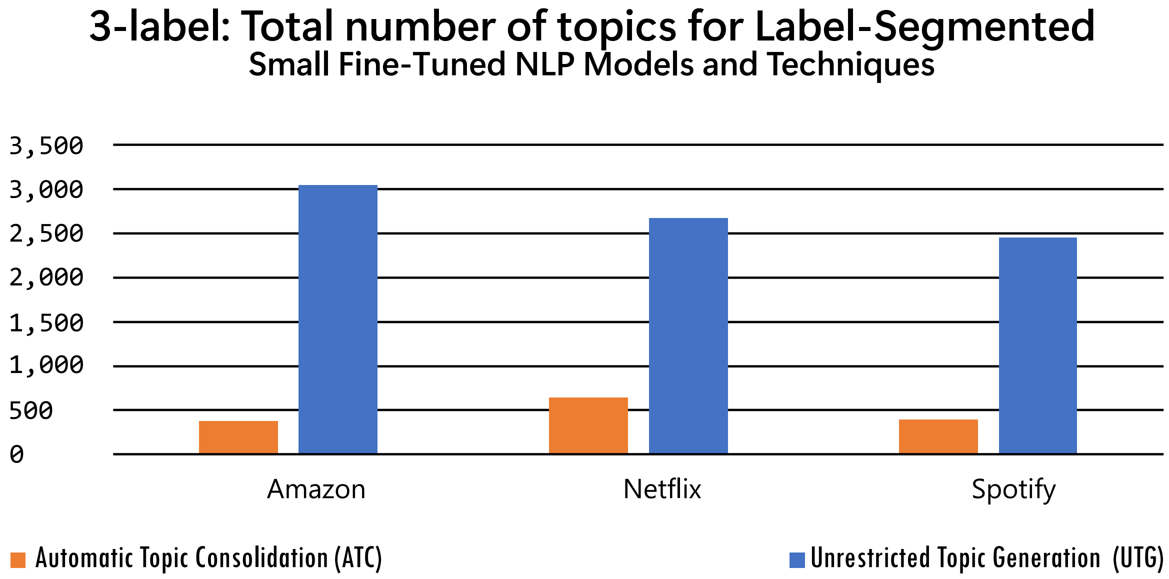


Figure 5.9: Bar Graph of the Total Number of Topics for All Datasets and Approaches in 3-Label Classification Using LS Topic Modeling.

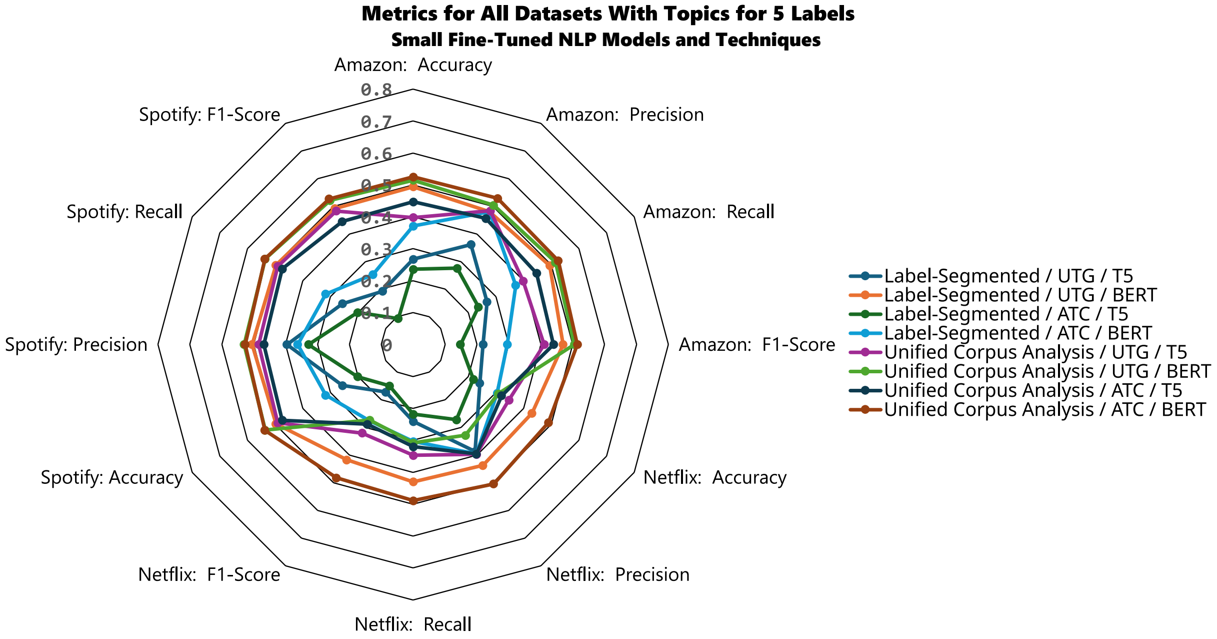


Figure 5.10: Radar Chart With Metrics for BERT and T5 Models Across Datasets for 5-Label Sentiment Analysis With Topics.

0.503, and 0.56, respectively. The overall average coherence for UCA is 0.546, while for LS, it is 0.668. Although the values for LS were higher, it did not reflect in an improvement in sentiment analysis.

Five-labels results

We then examined the five-class labeling scheme with topic modeling integration under both UTG and ATC scenarios, as presented in Tables 5.7 and 5.9, respectively. These findings shed light on the efficacy of leveraging topic modeling to enhance sentiment analysis across varying levels of sentiment granularity. Figure 5.10 shows these results where the lines close to the center are the ones regarding LS, and those far from it are regarding UCA.

In UCA (Table 5.7), the highest F1 score of 0.527 was achieved by the Spotify dataset using the BERT model under the ATC scenario, closely followed by 0.521 under UTG, this value was the same for the Amazon under ATC, under UTG the value for Amazon was 0.5. Additionally, the Netflix dataset scored 0.483 with BERT under ATC, a result similar (0.482) for Spotify under UTG with the model T5. However, under T5 and ATC, the results were notably poorer, with F1 scores of 0.445, 0.44, and 0.288 for the Spotify, Amazon, and Netflix datasets, respectively, the first result being the worst for Spotify. Under UTG, T5 scored 0.411 and 0.32 for Amazon and Netflix. The Netflix dataset also recorded the lowest score of 0.273 under UTG with BERT, highlighting the model's struggle in UCA. BERT had an average F1 score of 0.47, whereas T5's average was 0.398, making BERT the higher performer. The average F1 score for UTG was 0.419, while ATC had a higher average of 0.449. The overall average F1 score was 0.434.

Dataset	Approach	Accuracy	Precision	Recall	F1
Amazon	T5 / UTG	0.398	0.484	0.398	0.411
	BERT / UTG	0.515	0.504	0.515	0.505
	T5 / ATC	0.447	0.456	0.447	0.44
	BERT / ATC	0.525	0.529	0.525	0.515
Netflix	T5 / UTG	0.347	0.397	0.347	0.32
	BERT / UTG	0.308	0.328	0.308	0.273
	T5 / ATC	0.32	0.396	0.32	0.288
	BERT / ATC	0.489	0.504	0.489	0.482
Spotify	T5 / UTG	0.49	0.484	0.49	0.483
	BERT / UTG	0.536	0.53	0.536	0.521
	T5 / ATC	0.474	0.468	0.474	0.445
	BERT / ATC	0.537	0.527	0.537	0.527

Table 5.7: Evaluation of BERT and T5 Models in 5-Label Topic-Informed Sentiment Analysis with UTG and ATC Using UCA Across Datasets.

Dataset	Approach	Coherence	Topics
Amazon	UTG	0.569	2, 603
	ATC	0.569	1, 225
Netflix	UTG	0.457	2, 036
	ATC	0.457	409
Spotify	UTG	0.487	3, 911
	ATC	0.487	565

Table 5.8: Topic Coherence and Number of Topics in UCA for 5-Label Classification.

It was evaluated whether the addition of topics fails to create a significant difference in the F1 score between BERT and T5 in the five-label classification setting under UCA. The p-value produced a value of 0.094, so we failed to reject the null hypothesis, indicating no significant difference between the models. Similarly, comparing the UTG and ATC scenarios yielded a p-value of 0.687, again confirming no significant difference between these scenarios.

The coherence scores in this configuration are presented in Table 5.8 as a snapshot across different scenarios, as observed in Figure 5.11. Although UTG generated significantly more topics (2, 603, 2, 036, and 3, 911, respectively), ATC effectively consolidated these into 1, 225, 409, and 565 (Figure 5.12) topics while maintaining consistent topic coherence scores around 0.559 (Amazon), 0.457 (Netflix), and 0.457 (Spotify). ATC demonstrated the potential for impactful thematic structures through efficient topic consolidation while preserving topic quality.

Turning to the Label Segment scenario presented in Table 5.9, the BERT model under the UTG scenario yielded the best results for the Spotify, Amazon, and Netflix datasets, achieving F1 scores of 0.492, 0.47, and 0.417, respectively.

Conversely, the T5 model delivered the lowest outcome, with scores as low as 0.148 and 0.149 for the Amazon and Netflix datasets, respectively, and 0.094 for the Spotify dataset under ATC. On average, BERT achieved an F1 score of 0.367, while T5 had an

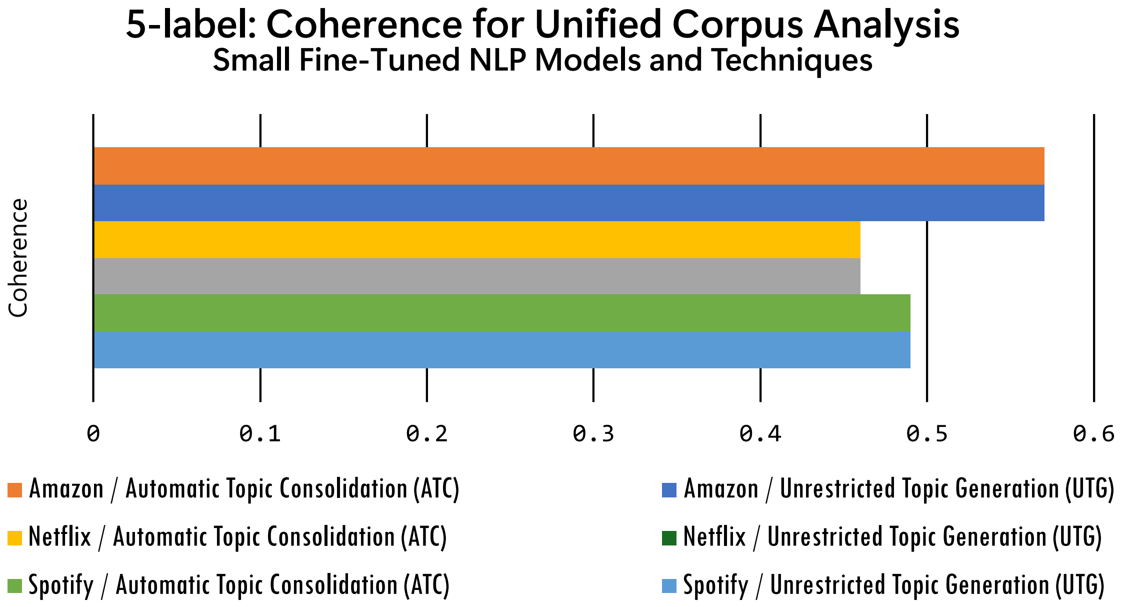


Figure 5.11: Bar Graph of Coherence Scores Across Datasets and Approaches for 5-Label Classification Using UCA.

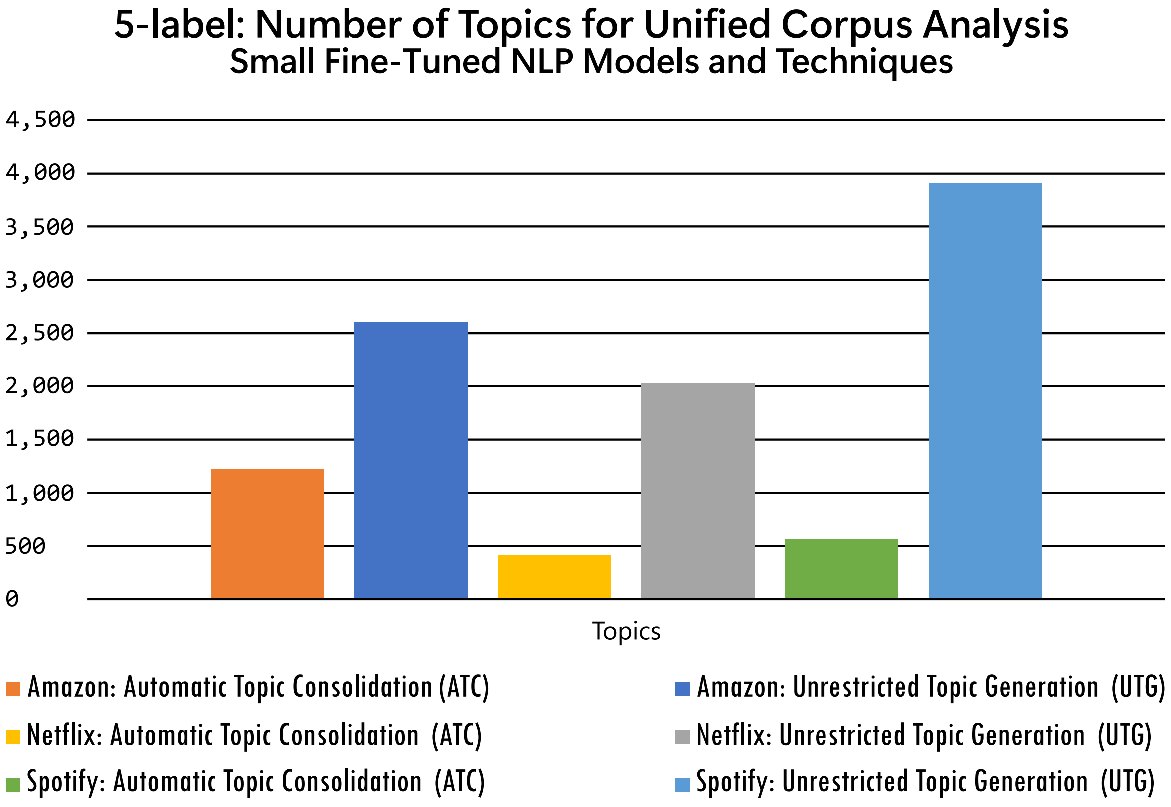


Figure 5.12: Bar Graph Showing the Number of Topics per Class for All Datasets and Approaches in 5-Label Classification Using UCA.

Dataset	Approach	Accuracy	Precision	Recall	F1
Amazon	T5 / UTG	0.267	0.362	0.267	0.219
	BERT / UTG	0.494	0.48	0.494	0.47
	T5 / ATC	0.236	0.276	0.236	0.148
	BERT / ATC	0.371	0.48	0.371	0.295
Netflix	T5 / UTG	0.241	0.387	0.241	0.172
	BERT / UTG	0.43	0.437	0.43	0.417
	T5 / ATC	0.218	0.272	0.218	0.149
	BERT / ATC	0.304	0.396	0.304	0.275
Spotify	T5 / UTG	0.256	0.396	0.256	0.193
	BERT / UTG	0.497	0.505	0.497	0.492
	T5 / ATC	0.201	0.328	0.201	0.094
	BERT / ATC	0.317	0.362	0.317	0.253

Table 5.9: Evaluation of BERT and T5 Models in 5-Label Topic-Informed Sentiment Analysis with UTG and ATC Using LS Across Datasets.

average of 0.162. The UTG scenario had a higher average F1 score of 0.327 compared to 0.202 for ATC. Overall, the average F1 score across all scenarios was 0.264.

The results of the Wilcoxon Signed-Rank Test indicate a statistically significant difference in F1 scores between BERT and T5 in the five-label classification setting under LS, with a p-value of 0.031, which is below the 0.05 significance level. This leads us to reject the null hypothesis that adding topics does not affect the F1 score difference between BERT and T5. Similarly, for the UTG and ATC scenarios, the test yielded a p-value of 0.031, leading us to reject the null hypothesis that adding topics does not significantly impact the F1 scores between these scenarios.

The topic coherence scores exhibit consistency across the UTG and ATC modeling approaches, with slight variations, as seen in Table 5.10 and in Figure 5.13. In the Amazon dataset, the coherence values are around 0.707, 0.693, 0.658, 0.657, and 0.69 for the “very negative”, negative, neutral, positive, and “very positive” classes, respectively. For the Netflix dataset, the coherence values are 0.755, 0.738, 0.722, 0.751, and 0.47 for the corresponding classes. For the Spotify dataset, the coherence scores are 0.654, 0.617, 0.706, 0.656, and 0.591. The ATC approach produced a more consolidated set of topics: 651 for Amazon (59, 146, 26, 151, 269 per class), 733 for Netflix (48, 77, 284, 113, 211 per class), and 854 for Spotify (521, 198, 96, 10, 29 per class). In contrast, the UTG method generated a larger number of topics: $2,719$ for Amazon (560, 517, 550, 541, 551 per class), $2,373$ for Netflix (428, 441, 475, 491, 538 per class), and $4,098$ for Spotify (827, 817, 851, 818, 785 per class). Figure 5.14 presents the number of topics per class. Figure 5.15 presents the total number of topics conveyed.

In this 5-label setting, the coherence values between ATC and LS remained unchanged, as we observed in the 3-label setting. Compared to UCA, the coherence values for LS are also higher. Specifically, the mean coherence values for LS are 0.681 for Amazon, 0.687 for Netflix, and 0.645 for Spotify, all of which surpass the corresponding UCA values of 0.569, 0.457, and 0.487, respectively. The overall average coherence for UCA is 0.504, while for LS it is 0.671. The same conclusion from the 3-label setting applies here: although LS

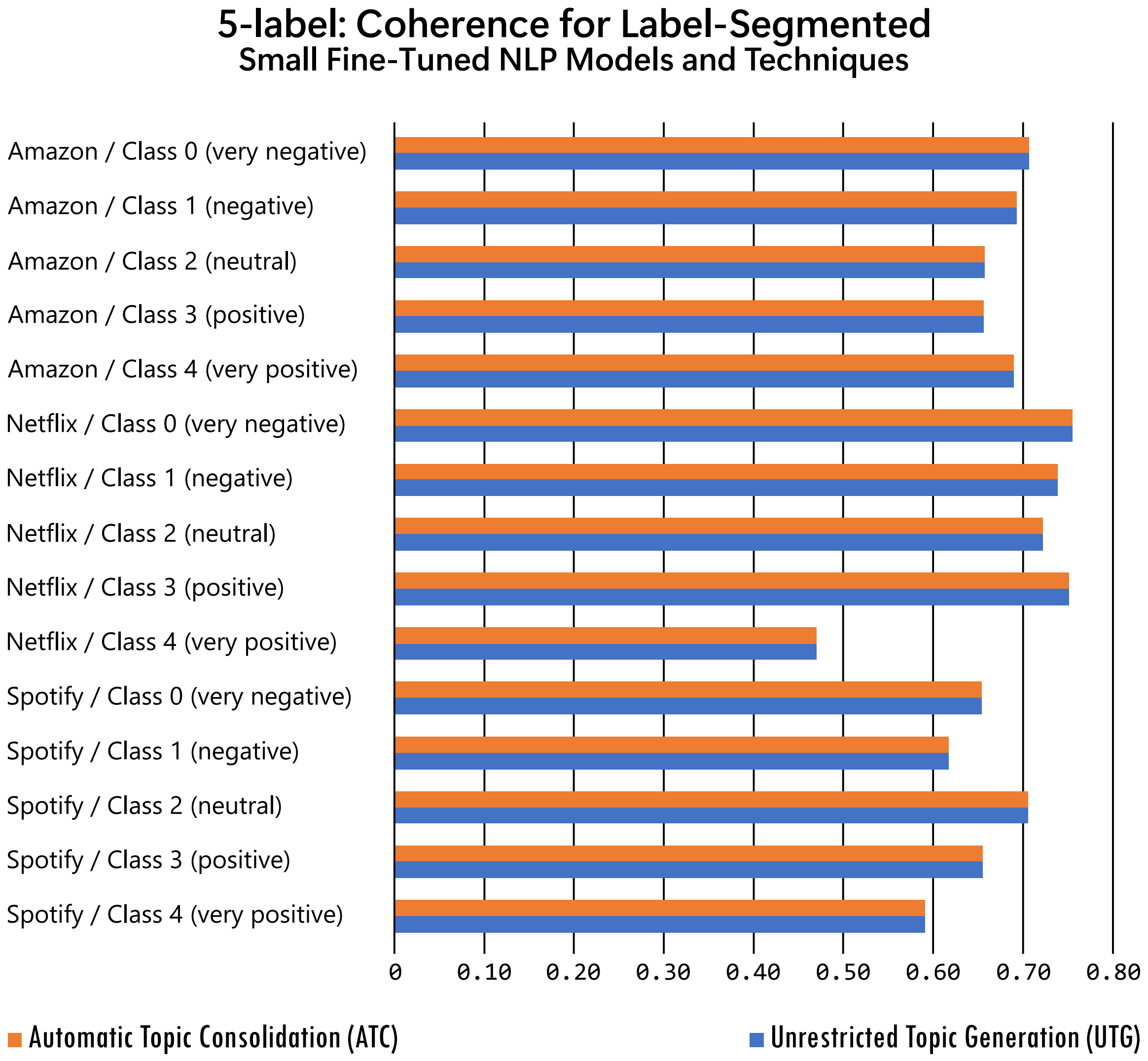


Figure 5.13: Bar Graph of Coherence Scores for All Datasets and Approaches in 5-Label Classification Using LS Topic Modeling.

Dataset	Label	UTG		ATC	
		Coherence	Topics	Coherence	Topics
Amazon	0 (very negative)	0.707	560	0.707	59
	1 (negative)	0.693	517	0.693	146
	2 (neutral)	0.658	550	0.658	26
	3 (positive)	0.657	541	0.657	151
	4 (very positive)	0.69	551	0.69	269
Netflix	0 (very negative)	0.755	428	0.755	48
	1 (negative)	0.738	441	0.738	77
	2 (neutral)	0.722	475	0.722	284
	3 (positive)	0.751	491	0.751	113
	4 (very positive)	0.47	538	0.47	211
Spotify	0 (very negative)	0.654	827	0.654	521
	1 (negative)	0.617	817	0.617	198
	2 (neutral)	0.706	851	0.706	96
	3 (positive)	0.656	818	0.656	10
	4 (very positive)	0.591	785	0.591	29

Table 5.10: Topic Coherence and Number of Topics in LS for 5-Label Classification.

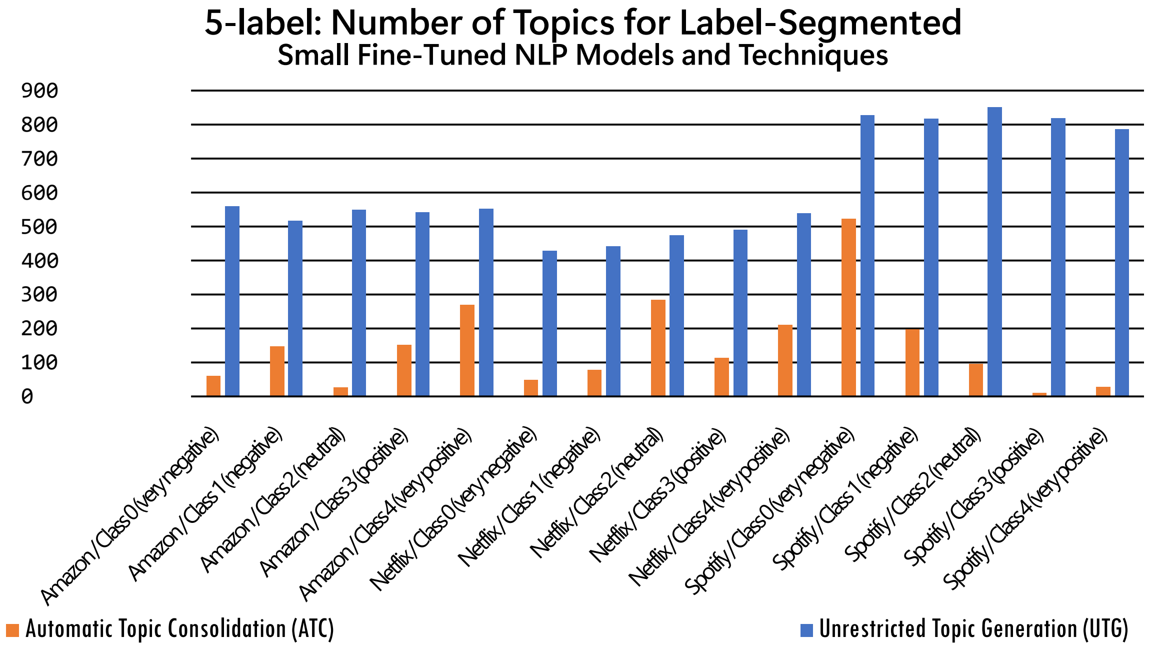


Figure 5.14: Bar Graph of the Number of Topics per Class for All Datasets and Approaches in 5-Label Classification Using LS Topic Modeling.

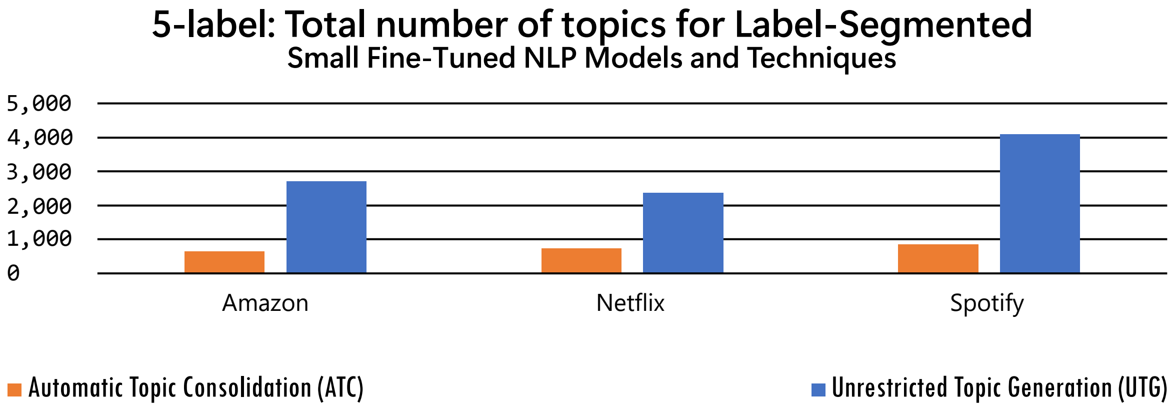


Figure 5.15: Bar Graph of the Total Number of Topics for All Datasets and Approaches in 5-Label Classification Using LS Topic Modeling.

shows higher coherence values, this does not translate into an improvement in sentiment analysis.

5.3 Discussion

We discuss the comparative evaluation of BERT and T5, assessing their potential with and without topics in various settings.

UCA and ATC emerge as more effective approaches for sentiment analysis compared to their respective counterparts, LS topic modeling and UTG. Across various datasets and models, UCA consistently demonstrated superior results, with higher F1 scores indicating a better balance between precision and recall. By consolidating topics across the entire corpus rather than segmenting them by labels, UCA provides a unified representation of topics, improving the model's ability to generalize across different sentiment labels. Statistical tests, such as the Wilcoxon signed-rank test, showed no significant difference in some cases between UCA and LS. However, UCA's better average results with fewer topics made it the more efficient and practical choice for sentiment analysis tasks.

Similarly, ATC consistently outperformed UTG in terms of average F1 scores, showing better overall accuracy, precision, and recall across various datasets and models. Wilcoxon Signed-Rank Tests revealed that while both UCA and LS showed no significant differences between ATC and UTG in three-label classifications, ATC demonstrated a significant advantage in five-label classifications using the LS approach. This indicates that ATC excels in more detailed topic labeling scenarios, making it the superior choice for sentiment analysis tasks by offering a more streamlined set of topics without sacrificing effectiveness.

When comparing language models, BERT consistently outperformed T5 across all datasets and labeling schemes. In the three-label classification without topics, BERT achieved higher F1 scores than T5 across all datasets, including Amazon, Netflix, and Spotify. For instance, BERT's F1 score for the Netflix dataset was 0.625, while T5's was 0.494. On average, BERT's F1 score across models was 0.635, compared to T5's 0.58. In the five-label classification, BERT continued to slightly outperform T5, particularly in datasets like Spotify, where BERT achieved an F1 score of 0.522 compared to T5's 0.509.

Even when incorporating topic-informed sentiment analysis, BERT maintained stronger outcomes in both UTG and ATC scenarios.

In conclusion, both UCA and ATC stand out as superior methods for sentiment analysis due to their ability to produce more focused, generalizable topics and better overall model results. BERT demonstrates consistent superiority over T5, making it the better model across both three-label and five-label settings and in topic-informed analyses. This combination of streamlined topic modeling and high-performing language models establishes a strong foundation for effective sentiment analysis. The coherence values didn't prove itself relevant enough to make an impact in sentiment analysis.

In the analysis of sentiment classification without topic modeling, the three-label scheme consistently showed better results than the five-label scheme. The average F1 score for the three-label scheme was 0.607, while the five-label scheme scored 0.504. A Wilcoxon Signed-Rank Test confirmed a statistically significant difference, yielding a p-value of 0.031, which indicates the three-label scheme's superior effectiveness in this scenario.

When topic modeling was introduced, specifically through the UCA technique, the values declined for both labeling schemes; however, the three-label scheme still achieved better results. The average F1 scores were 0.592 for the three-label scheme and 0.434 for the five-label scheme. The Wilcoxon Signed-Rank Test indicated a highly significant p-value of 4.9×10^{-4}, reinforcing the conclusion that the three-label scheme remains more effective even with the addition of topic modeling.

Additionally, when applying topic modeling via label segmentation, the three-label scheme again outperformed the five-label scheme. The average F1 scores were 0.387 for the three-label scenario compared to 0.264 for the five-label scenario. The Wilcoxon Signed-Rank Test once more demonstrated a highly significant difference, with a p-value of 4.9×10^{-4}, further confirming the superior results of the three-label scheme.

Sentiment analysis without topics yields an average F1 score of 0.555 – 0.607 for the three-label scenario and 0.504 for the five-label scenario – indicating better outcomes without topic integration. This may be because topic modeling introduces additional complexity that doesn't always align with sentiment labels, particularly in finer-grained, five-label tasks. When topics are included[7], the average F1 score decreases to 0.513 (0.592 for three-label and 0.434 for five-label scenarios), suggesting that topic integration may slightly dilute sentiment-specific cues, especially in the more granular five-label setting, thereby reducing F1 scores.

5.4 Final Remarks

BERT and T5 were compared for sentiment analysis, incorporating various topic modeling techniques. BERT consistently outperformed T5, especially in three-class sentiment classification tasks. UCA and ATC emerged as the most effective topic modeling methods, producing streamlined topics that improved model results. Three-label classification consistently yielded better results than five-label classification, both with and without

[7] Under UCA.

topic modeling. In some cases, particularly for more detailed sentiment classifications, the inclusion of topic modeling slightly reduced results. Our findings indicate that combining streamlined topic modeling techniques with language models like BERT and T5 does not enhance sentiment analysis, especially in more nuanced sentiment classifications.

As we transition to Pre-Trained Large Language Models (Chapter 6), the focus shifts to a more scalable and versatile solution for similar tasks. Unlike small fine-tuned models, pre-trained large language models, such as Meta Llama 3 8B and Mixtral 8x7B, may offer the potential to handle much larger contexts and more diverse tasks. Chapter 6 will explore applying these advanced models in the same domain – topic modeling and sentiment analysis. The pre-trained models will be examined under various learning paradigms (Zero-Shot, One-Shot, and Few-Shot), offering a broader comparison between small fine-tuned techniques and more generalized pre-trained architectures.

This bridge between our two evaluated approaches in this M.Sc. dissertation sets the stage for comparing and contrasting small fine-tuned models and techniques versus pre-trained large language models, providing insights into their strengths and limitations, particularly in real-world sentiment and topic analysis applications.

Chapter 6

Evaluating Pre-Trained Large Language Models

The primary objective of this chapter is to evaluate the effectiveness of pre-trained large language models for sentiment analysis, both with and without the integration of topic modeling. By comparing model configurations and learning approaches, including zero-shot, one-shot, and few-shot learning, we aim to establish insights into their results across various sentiment classification tasks. This evaluation highlights the advantages and limitations of each configuration, providing a comprehensive analysis of model suitability for sentiment analysis tasks of differing complexities.

This chapter is organized as follows: Section 6.1 presents the methodological steps involved in configuring and fine-tuning the pre-trained large language models. This section also explains the configurations explored in this study and the evaluation pipeline. In Section 6.2 it examines the outcomes of each model with and without the integration of topics, covering three- and five-label classification scenarios. The Discussion section (cf. Section 6.3) provides a comparative analysis of the models, discussing insights derived from both topic-influenced and non-topic scenarios. Then the Final Remarks (cf. Section 6.4) summarizes the key findings from this chapter and introduces the transition to Chapter 7, which delves into the broader comparison between small fine-tuned models and pre-trained large language models.

6.1 Specific Procedures for Topic Modeling and Sentiment Analysis Tasks

In Figure 6.1 a visual representation of the different settings explored for the Pre-Trained Large Language Models is presented. This will be further explained in the following subsections inside this one.

This evaluation includes statistical tests to verify the significance of differences observed between models. The following hypotheses were tested using the Wilcoxon signed-rank test:

- **Three-label classification without topics**: There is no significant difference

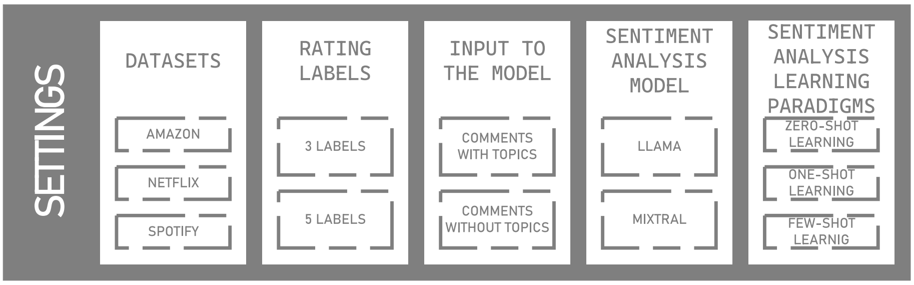


Figure 6.1: Different Configurations Explored For The Pre-Trained Large Language Models.

in the F1 scores between Meta Llama 3 8B and Mixtral 8x7B for the three-label classification without topics.

- **Five-label classification without topics**: In the absence of topics, there is no significant difference in the F1 scores between Meta Llama 3 8B and Mixtral 8x7B for the five-label classification.
- **Three-label classification with topics**: Adding topics doesn't bring about a significant difference in F1 scores between Meta Llama 3 8B and Mixtral 8x7B for the three-label classification setting.
- **Five-label classification with topics**: Adding topics fails to yield a significant difference in F1 scores between the Meta Llama 3 8B and Mixtral 8x7B for the five-label classification.

The use of the Wilcoxon signed-rank test throughout these comparisons allowed us to validate our observations rigorously, ensuring that the results reflect consistent model behavior across various sentiment analysis tasks.

In this phase of the research, we leverage GroqCloud[1], a platform that offers a serverless API specifically designed for AI inference tasks. GroqCloud provides developers with a streamlined solution to integrate advanced AI capabilities into their applications without managing complex infrastructure.

To utilize GroqCloud, developers first obtain an API key, which serves as their authentication credential. This key allows them to make secure inference requests to the GroqCloud servers, where the AI models are hosted and optimized for efficiency.

GroqCloud is particularly noteworthy for its support of state-of-the-art models, especially those optimized for NLP tasks. Our study specifically employed two leading models: Meta Llama 3 8B known for its efficiency and high outcomes with a context window of 8,192 tokens, and Mixtral 8x7B, recognized for its ability to process and understand longer text sequences with an expansive context window of 32,768 tokens.

6.1.1 Topic Modeling

The process involves using LLMs to extract topics (or keyphrases) from user reviews, which are then clustered and merged to form a concise summary of core themes. Figure

[1] Groq website. `https://groq.com`. Accessed on July 29, 2024

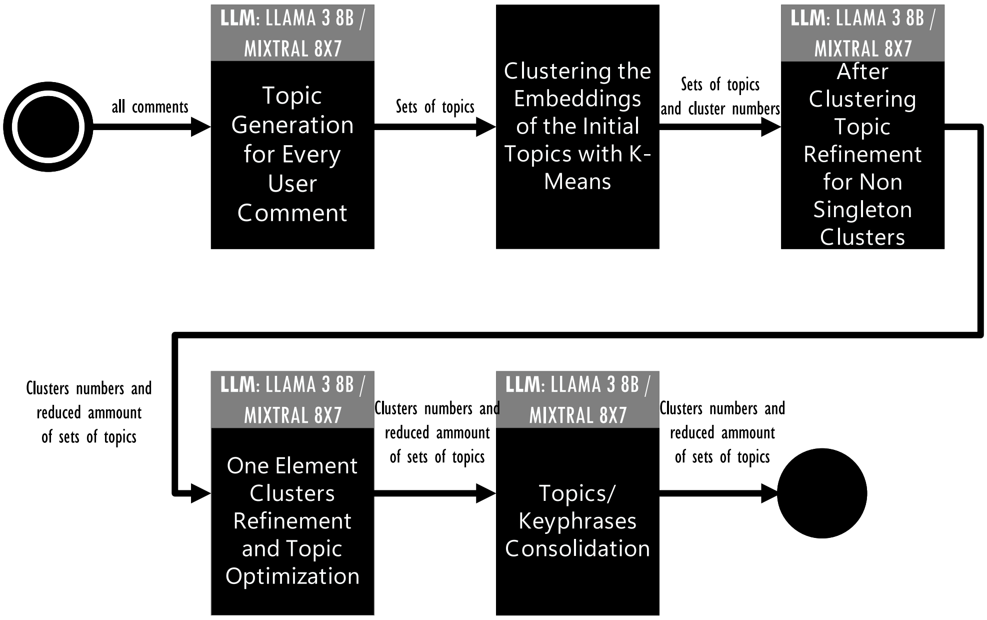

all comments
LLM: LLAMA 3 8B / MIXTRAL 8X7
Topic Generation for Every User Comment
Sets of topics
Clustering the Embeddings of the Initial Topics with K-Means
Sets of topics and cluster numbers
LLM: LLAMA 3 8B / MIXTRAL 8X7
After Clustering Topic Refinement for Non Singleton Clusters
Clusters numbers and reduced ammount of sets of topics
LLM: LLAMA 3 8B / MIXTRAL 8X7
One Element Clusters Refinement and Topic Optimization
Clusters numbers and reduced ammount of sets of topics
LLM: LLAMA 3 8B / MIXTRAL 8X7
Topics/ Keyphrases Consolidation
Clusters numbers and reduced ammount of sets of topics

Figure 6.2: Pipeline to conduct topic modeling using Pre-Trained Large Language Models.

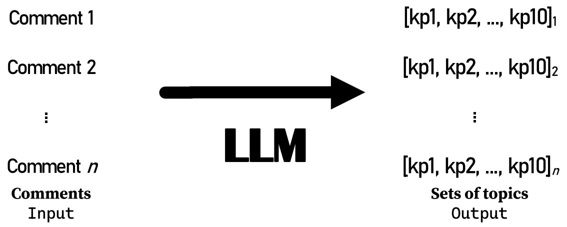


Figure 6.3: Topics extraction from the comments. For every comment, there is an associated set of topics.

6.7 visualizes this pipeline, which will be detailed further.

The initial step involves a Python script that extracts topics for each user comment by sending requests to the GroqCloud API, which returns a list of ten topics per comment. This initial step is referred to as “Topic Generation for Every User Comment,” and in Figure 6.3 shows the desired transformation from the comments. Below in Prompt 1 there is the instruction given to the LLM for this initial step. These generated topics are then collected and appended to the original dataset as a new column. To ensure the topics are formatted correctly, the script includes a function that extracts and properly formats the relevant part of the API response. Once all the topics have been generated and added to the dataset, the new dataset is saved back to a Comma-Separated Values (CSV) file, ready for further analysis. In short, the extracted keyphrases are added as a new column to the dataset and saved as a CSV file.

The process, titled “Clustering the Embeddings of the Initial Topics with *K-Means* ,” continues with loading and tokenizing the data using a pre-trained BERT model (bert-base-uncased). This model converts the keyphrases into tokens, from which word and

Prompt 1 Topic Generation for Every User Comment Prompt.

```
I am analyzing user comments from the Google Play Store about the Amazon
 ↪  App to determine the topics being discussed. For each comment,
 ↪  generate a comprehensive set of topics/keyphrases that accurately
 ↪  capture the topics.
Text: "{text}"

Please just provide the output (keyphrases) as a list of strings in the
 ↪  following format: ["keyphrase0", " keyphrase1", "keyphrase2",
 ↪  "keyphrase3", "keyphrase4", "keyphrase5", "keyphrase6", "keyphrase7",
 ↪  "keyphrase8", "keyphrase9"].
Ensure that the set contains exactly 10 keyphrases. If the set has less
 ↪  than 10 keyphrases, please add new ones to reach a total of 10. If
 ↪  the set has more than 10 keyphrases, please identify the most
 ↪  important ones and remove any redundant or irrelevant keyphrases to
 ↪  reach a total of 10. If the set already contains exactly 10
 ↪  keyphrases, please leave it as is.
Don't provide anything more than the set of keyphrases.
```

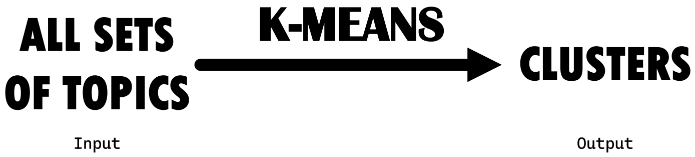


Figure 6.4: Simplified visualization of Clustering the Embeddings of the Initial Topics with *K-Means* process.

sentence embeddings[2] are generated. These embeddings are then reshaped into a 2D array and clustered using *K-Means* , grouping the keyphrases into 500 clusters. In Figure 6.4, there is a simplification in a visual way of this process where the clusterization is conducted with *K-Means* from the sets of topics. The output is another CSV file containing a new column with the cluster numbers.

Once the clustering is complete, the script goes a step further by refining the clusters, this step is called “After Clustering Topic Refinement for Non Singleton Clusters”. It consolidates topics within each cluster that contain more than one set of topics, generating a more concise (with ten topics) and representative set of topics that capture the main themes discussed in the user comments. This is achieved by sending the grouped topics back to the LLM, which then returns a refined set of topics, as showcased in Figure 6.5. The Prompt 2 was used in this step, in the variable named ‘texto’ is provided the sets to be summarized.

[2] Word embeddings represent individual words, while sentence embeddings capture the meaning of entire comments.

Prompt 2 After Clustering Topic Refinement for Non Singleton Clusters Prompt.

```
Please generate a single set of topics/keyphrases that combines the main
 ↪  theme and the specific issues from the following sets of information:
{texto}
The output should be a single set of topics/keyphrases that captures the
 ↪  main theme and the specific issues from both sets in the following
 ↪  format: ["keyphrase0", "keyphrase1", "keyphrase2", ...]
Ensure that the set contains exactly 10 topics/keyphrases. If the set has
 ↪  less than 10 topics/keyphrases, please summarize the existing
 ↪  keyphrases and add new ones to reach a total of 10. If the set has
 ↪  more than 10 topics/keyphrases, please identify the most important
 ↪  ones and remove any redundant or irrelevant topics/keyphrases to
 ↪  reach a total of 10. If the set already contains exactly 10
 ↪  topics/keyphrases, please leave it as is.
Don't provide anything more than the set of topics/keyphrases.
```

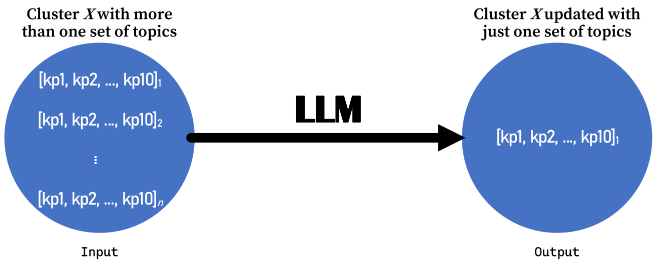

Cluster X with more than one set of topics
[kp1, kp2, ..., kp10]$_1$
[kp1, kp2, ..., kp10]$_2$
⋮
[kp1, kp2, ..., kp10]$_n$
Input
LLM
Cluster X updated with just one set of topics
[kp1, kp2, ..., kp10]$_1$
Output

Figure 6.5: After Clustering Topic Refinement for Non Singleton Clusters image representation.

For clusters that contain only a single set of keyphrases, the clustering process is further refined by sending these keyphrases back to the LLM for an additional round of merging, a step known as “One Element Clusters Refinement”. During this process, the LLM evaluates whether the set of topics from one cluster shares common themes with sets of topics from other clusters with just one element. If commonalities are found, the clusters are merged; if not, the cluster remains unchanged, ensuring that the output ends up with exactly ten topics (one set), with any redundant or less relevant keyphrases removed, resulting in a concise and meaningful set, in Figure 6.6 there is a visual explanation. The Prompt 3 is the one used for that end:

Prompt 3 One Element Clusters Refinement and Topics/Keyphrases Consolidation Prompt.

```
Please perform the following tasks:

Cluster the given sets of topics/keyphrases into groups based on their
 ↪  themes. Indicate the number of the set that contributes to each
 ↪  cluster. The number of clusters is not predefined, so please create
 ↪  as many as you see fit based on the data.
For clusters that contain more than one set of topics/keyphrases, merge
 ↪  all the sets to generate one that synthesizes all of them. The goal
 ↪  is to represent the overall theme of the cluster in a single set of
 ↪  topics/keyphrases. If a cluster has only one set of
 ↪  topics/keyphrases, leave it as is.
Then, list the sets that contribute to the cluster and the merged
 ↪  topics/keyphrases.

Here are the sets of topics/keyphrases to work with: "{texto}"

Ensure that the set contains exactly 10 topics/keyphrases. If the set has
 ↪  less than 10 topics/keyphrases, please summarize the existing
 ↪  topics/keyphrases and add new ones to reach a total of 10. If the set
 ↪  has more than 10 topics/keyphrases, please identify the most
 ↪  important ones and remove any redundant or irrelevant
 ↪  topics/keyphrases to reach a total of 10. If the set already contains
 ↪  exactly 10 topics/keyphrases, please leave it as is.
Don't provide nothing more than the set of topics/keyphrases.
The output should be presented in the following format:

Sets [list of set numbers]
Merged keyphrases: ['keyphrase 1', 'keyphrase 2', 'keyphrase 3', ...]
```

This same prompt is used again to check other commonalities with all the resulting new topics (Figure 6.7), this final step is called “Topics/Keyphrases Consolidation.”

The described process automates the extraction of meaningful topics from user comments and demonstrates the effectiveness of utilizing powerful pre-trained large language models through a cloud-based platform like GroqCloud. This approach is integral to the

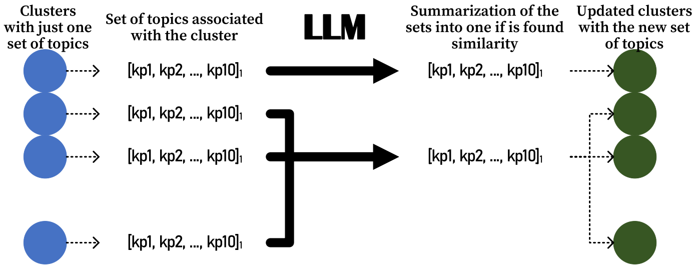

Clusters with just one set of topics
Set of topics associated with the cluster
LLM
Summarization of the sets into one if is found similarity
Updated clusters with the new set of topics
[kp1, kp2, ..., kp10]1
[kp1, kp2, ..., kp10]1
[kp1, kp2, ..., kp10]1
[kp1, kp2, ..., kp10]1
[kp1, kp2, ..., kp10]1
[kp1, kp2, ..., kp10]1

Figure 6.6: One Element Clusters Refinement, the LLM merges them based on common themes to ensure a concise set of exactly ten keyphrases.

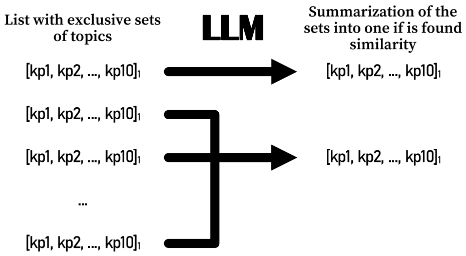


Figure 6.7: Topics/Keyphrases Consolidation, where the final sets of topics is produced.

study's methodology, particularly in demonstrating the application of LLMs in generating insightful topics from customer reviews.

6.1.2 Sentiment Analysis

We describe the method for sentiment analysis by covering scenarios with and without topic consideration, as well as three and five sentiment label classifications. It employs pre-trained large language models and evaluates the efficacy of prompts using zero-shot, one-shot, and few-shot learning approaches. In the following, we present how each approach was applied:

- **Zero-Shot Learning**: This approach involved using the pre-trained LLMs directly without any additional training on the specific dataset. The models, equipped with their inherent understanding of language, were able to predict the sentiment of the comments based solely on their pre-existing knowledge.

- **One-Shot Learning**: In this approach, a single example for every class from the dataset was provided to the models, allowing them to adjust their predictions based on these references. This method is advantageous when only a minimal amount of labeled data is available.

- **Few-Shot Learning**: Here, the models were given three – for the three label setting – or five – in case of the five label setting – examples from each dataset class. This sample helped the models better understand the context and nuances of the comments, leading to more accurate sentiment predictions.

All prompts and code used in the sentiment analysis process are available in the GitHub repository[3]. This includes detailed implementations, sample prompts, and the full codebase necessary for replicating the analysis. The repository is organized to facilitate easy access and understanding, providing resources for anyone interested in exploring or extending this work.

6.2 Results

This section presents sentiment analysis results using Meta Llama 3 8B and Mixtral 8x7B pre-trained large language models. The experiments encompassed various learning conditions, including zero-shot, one-shot, and few-shot learning. Our analysis aims to thoroughly understand how these transformer-based models yield in sentiment classification tasks, highlighting their strengths and limitations under different conditions.

6.2.1 Results of Sentiment Analysis Without Topics

Three-labels results

In the three-class labeling scenario without the aid of topic modeling, the Meta Llama 3 8B model results varied across different datasets as seen in Table 6.1 and Figure 6.8. For the Spotify dataset, it achieved an F1 score of 0.672 in both Zero Shot and Few Shot scenarios, which suffered a decrease to 0.664 in One Shot. The Amazon dataset showed lower overall scores, with the model attaining an F1 score of 0.568 in Zero Shot, declining to 0.557 in One Shot, and further dropping to 0.55 in Few Shot. Similarly, for the Netflix dataset, the model scored 0.562 in Zero Shot, decreased to 0.525 in One Shot, and somewhat improved to 0.531 in Few Shot.

Interestingly, the model demonstrated peak outcomes in Zero-Shot scenarios, operating without prior examples. Its effectiveness decreased with One Shot and Few Shot learning, where providing examples and labels had a slightly detrimental effect. This pattern highlights that the model performed best in Zero Shot scenarios across all datasets, with results declining as more examples were provided. This unexpected outcome suggests that the Meta Llama 3 8B model may be more effective when making predictions without specific examples rather than with additional context through example-based learning.

Conversely, the Mixtral 8x7B model (cf. Table 6.1 and Figure 6.8) exhibit a contrasting behavior compared to the previously discussed model. In this case, providing more examples enhances the model's outcomes.

This trend is consistent across all three datasets examined. For the Spotify dataset, Mixtral 8x7B achieved its highest F1 score of 0.651 in the Few Shot learning scenario; this was followed by a score of 0.597 in One Shot learning, while Zero Shot learning yielded the lowest score of 0.509. The Amazon dataset showed a similar pattern; the model's results peaked in Few Shot learning with an F1 score of 0.547, One Shot learning

[3] H-IAAC/meta-7/tree/patrick-Sa_Tp/src/patrick/thesisExperimentsPrompts repository (GitHub). `https://github.com/H-IAAC/meta-7/tree/patrick-Sa_Tp/src/patrick/thesisExperimentsPrompts`. Accessed on August 12, 2024.

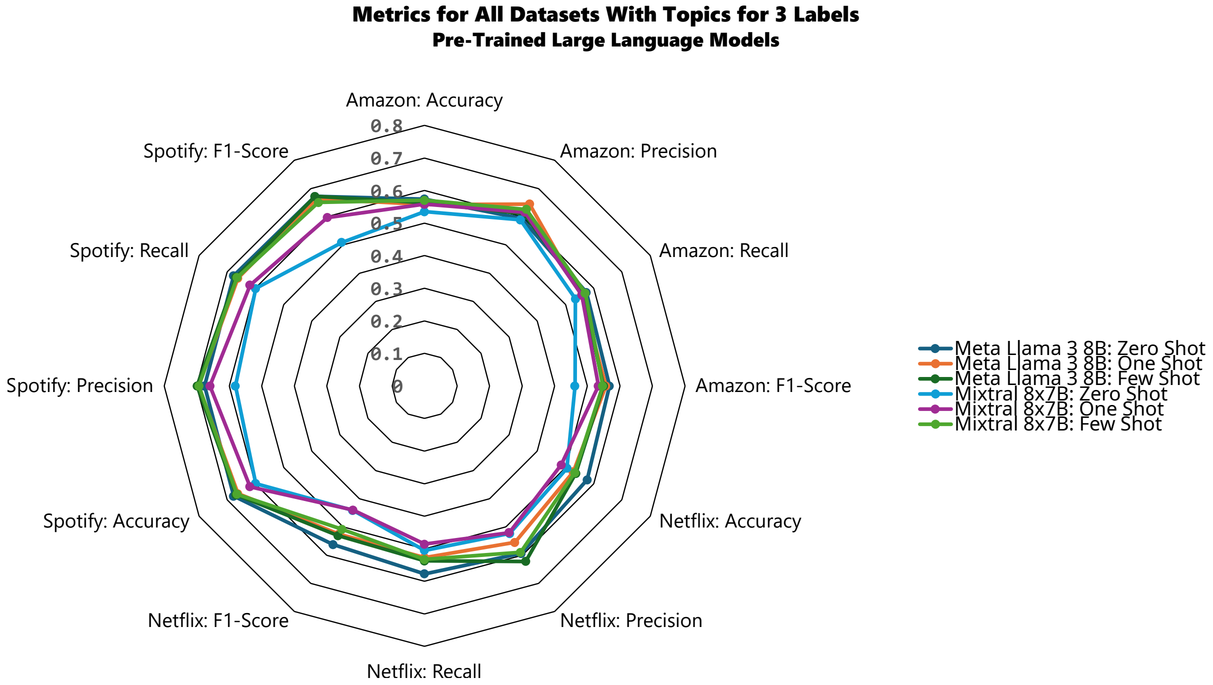


Figure 6.8: Radar Chart With Metrics for Meta Llama 3 8B and Mixtral 8x7B Models Across Datasets for 3-Label Sentiment Analysis Without Topics.

resulted in a slightly lower score of 0.534, while Zero Shot learning lagged with a score of 0.462. The Netflix dataset also aligned with this trend, Few Shot learning again produced the best results, with an F1 score of 0.508; for this dataset, Zero-Shot learning slightly outperformed One-Shot learning, with scores of 0.441 and 0.44, respectively, though the difference is minimal.

These findings indicate that Mixtral 8x7B benefits from exposure to examples, with its results generally improving as more examples are provided. This behavior stands in stark contrast to that observed in the previously discussed model, highlighting the varied responses different models can have to incremental learning scenarios.

When comparing the overall effectiveness, Meta Llama 3 8B achieved an average F1 score of 0.59, with specific scores of 0.60 for Zero Shot, 0.589 for One Shot, and 0.584 for Few Shot learning. On the other hand, Mixtral 8x7B had an average F1 score of 0.521, with scores of 0.471 in Zero Shot, 0.524 in One Shot, and 0.569 in Few Shot learning. These results indicate that, on average, Meta Llama 3 8B outperforms Mixtral 8x7B across all learning types.

The null hypothesis, stating that there is no significant difference in the F1 scores between Meta Llama 3 8B and Mixtral 8x7B for three-label classification without topics, was rejected based on the results of the Wilcoxon signed-rank test. The test yielded a p-value of 0.004. Given that this p-value is below the standard significance threshold of 0.05, we conclude that there is a statistically significant difference in the F1 scores between the two models.

Dataset	Metrics	Meta Llama 3 8B			Mixtral 8x7B		
		Zero Shot	One Shot	Few Shot	Zero Shot	One Shot	Few Shot
Amazon	Accuracy	0.574	0.557	0.563	0.535	0.558	0.57
	Precision	0.594	0.645	0.611	0.589	0.616	0.626
	Recall	0.574	0.557	0.563	0.535	0.558	0.57
	F1	0.568	0.557	0.55	0.462	0.534	0.547
Netflix	Accuracy	0.577	0.527	0.537	0.505	0.485	0.533
	Precision	0.594	0.555	0.622	0.523	0.52	0.59
	Recall	0.577	0.527	0.537	0.505	0.486	0.533
	F1	0.562	0.525	0.531	0.441	0.44	0.508
Spotify	Accuracy	0.677	0.662	0.668	0.599	0.619	0.665
	Precision	0.676	0.694	0.698	0.581	0.659	0.693
	Recall	0.677	0.662	0.668	0.599	0.619	0.665
	F1	0.672	0.664	0.672	0.509	0.597	0.651

Table 6.1: Evaluation Metrics for Meta Llama 3 8B and Mixtral 8x7B Models in 3-Label Sentiment Analysis Without Topics Across Datasets.

Five-labels results

In the five-label setting, the increase in the number of labels introduces more complexity with some notable differences in F1 scores across the datasets.

Metrics for the Meta Llama 3 8B model, outlined in Table 6.2 and visually represented in Figure 6.9, indicate variability across the different datasets.

For the Amazon dataset, the model achieved its highest F1 score of 0.307 in the One Shot learning scenario, followed closely by the Few Shot scenario with an F1 score of 0.306, the Zero Shot scenario yielded the lowest result for this dataset, with an F1 score of 0.23. In the case of the Netflix dataset, the model performed best in the Few Shot scenario with an F1 score of 0.293, followed by One Shot at 0.286. The Zero Shot scenario, however, resulted in the lowest F1 score of 0.187, which was also the lowest among all datasets evaluated. For the Spotify dataset, the highest F1 score was observed in the One Shot scenario at 0.313, with Few Shot and Zero Shot scenarios producing F1 scores of 0.263 and 0.245, respectively.

We observed a similar trend for the Mixtral 8x7B model (cf. Table 6.2). The highest F1 score for this model was achieved on the Spotify dataset in the Few Shot scenario, with a score of 0.402, followed by Zero Shot with 0.381, and One Shot with 0.339. For the Amazon dataset, the model performed best in the Zero Shot scenario with an F1 score of 0.384, followed by Few Shot at 0.369, and One Shot at 0.33. In the case of the Netflix dataset, the highest F1 score was observed in the One Shot scenario at 0.358, followed by Few Shot with 0.345 and Zero Shot with 0.335.

These results for both models demonstrated the complexity of results across different datasets and learning scenarios. While some consistencies exist, such as the general improvement from zero-shot to few-shot scenarios, notable variations exist.

For instance, the Mixtral 8x7B model shows a different pattern for each dataset, with Spotify performing best in Few Shot, Amazon in Zero-Shot, and Netflix in One Shot.

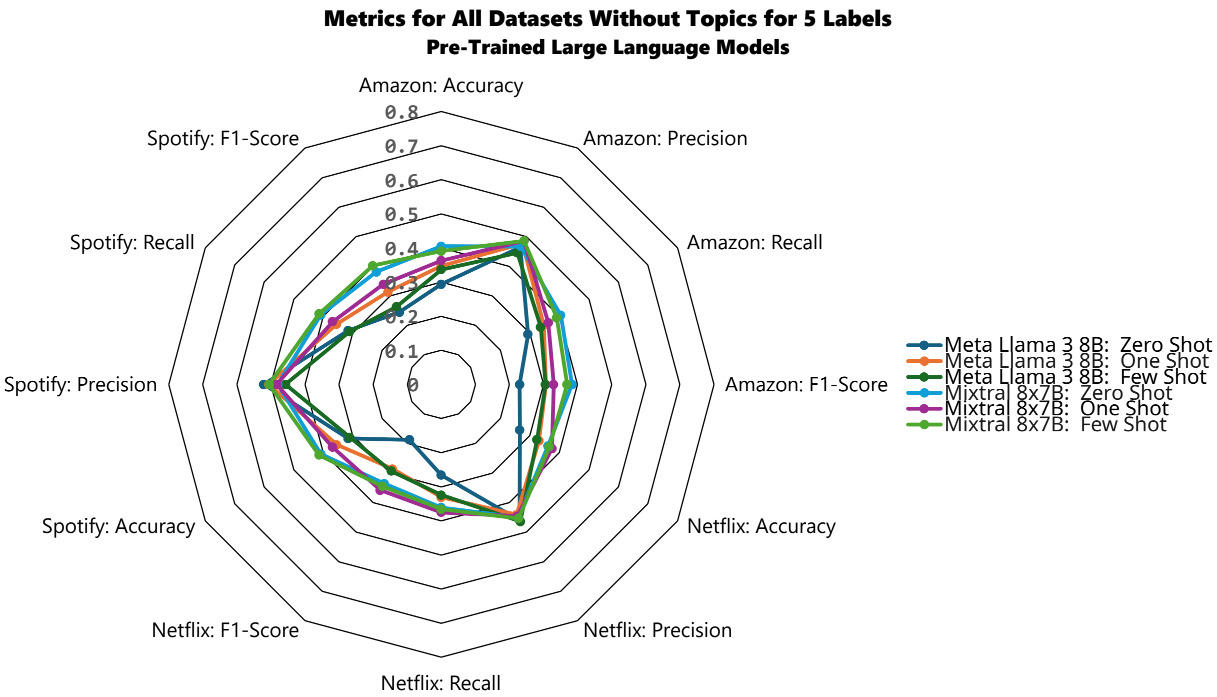


Figure 6.9: Radar Chart With Metrics for Meta Llama 3 8B and Mixtral 8x7B Models Across Datasets for 5-Label Sentiment Analysis Without Topics.

Dataset	**Metrics**	**Meta Llama 3 8B**			**Mixtral 8x7B**		
		Zero Shot	**One Shot**	**Few Shot**	**Zero Shot**	**One Shot**	**Few Shot**
Amazon	Accuracy	0.294	0.347	0.337	0.405	0.363	0.391
	Precision	0.464	0.477	0.446	0.467	0.486	0.486
	Recall	0.294	0.347	0.337	0.405	0.363	0.391
	F1	0.23	0.307	0.306	0.384	0.33	0.369
Netflix	Accuracy	0.266	0.331	0.324	0.361	0.375	0.366
	Precision	0.46	0.441	0.464	0.45	0.446	0.453
	Recall	0.266	0.331	0.324	0.361	0.375	0.366
	F1	0.187	0.286	0.293	0.335	0.358	0.345
Spotify	Accuracy	0.316	0.354	0.31	0.408	0.368	0.414
	Precision	0.521	0.506	0.457	0.48	0.483	0.505
	Recall	0.316	0.354	0.31	0.408	0.368	0.414
	F1	0.245	0.313	0.263	0.381	0.339	0.402

Table 6.2: Evaluation Metrics for Meta Llama 3 8B and Mixtral 8x7B Models in 5-Label Sentiment Analysis Without Topics Across Datasets.

This variability underscores the importance of considering dataset-specific characteristics when evaluating model efficiency. It also highlights the need for careful selection of learning scenarios based on the task and dataset. Furthermore, these findings suggest that while providing examples can often enhance model results, the optimal number of examples may differ depending on the model architecture and the nature of the data being processed. This complexity in model behavior across different scenarios emphasizes the need for comprehensive testing and evaluation when deploying these models in real-world applications.

On average, Meta Llama 3 8B has an F1 score of 0.27. Specifically, in Zero-Shot Learning, Meta Llama 3 8B has a mean score of 0.221; in One Shot Learning, it is 0.302; and in Few Shot Learning, it is 0.287. In contrast, Mixtral 8x7B has an average F1 score of 0.36. For Mixtral 8x7B, the scores are 0.367 in Zero Shot Learning, 0.342 in One Shot Learning, and 0.372 in Few Shot Learning. This demonstrates that Mixtral 8x7B consistently outperformed Meta Llama 3 8B across all learning types.

The Wilcoxon signed-rank test rejected the null hypothesis, which posited no significant difference in F1 scores between Meta Llama 3 8B and Mixtral 8x7B in the absence of topics. The test yielded a p-value of 0.004. Since this p-value is below the standard significance level of 0.05, it indicates a statistically significant difference in F1 scores between Meta Llama 3 8B and Mixtral 8x7B.

6.2.2 Results of Sentiment Analysis With Topics

Incorporating topic modeling into sentiment analysis can enhance model outcomes. However, the improvements are not uniform. While some scenarios see a slight boost, others[4] may experience a decrease in F1 scores. Thus, the effectiveness of topic modeling varies, and its impact should be evaluated carefully in different contexts. This section presents the results for both the Meta Llama 3 8B and Mixtral 8x7B models when topics are included in the sentiment classification process.

Three-labels results

The coherence scores (cf. Figure 6.10) detailed in Table 6.3 provide insights into the quality of topics generated and their relevance to sentiment classification tasks, taking into account both coherence and the number of topics (Figure 6.11) identified for each dataset.

The Meta Llama 3 8B model demonstrated strong coherence across all datasets, with scores of 0.988, 0.99, and 0.986 for Amazon, Netflix, and Spotify, respectively, while identifying 379 topics for Amazon, 401 for Netflix, and 356 for Spotify. Similarly, the Mixtral 8x7B model also achieved strong coherence, reaching a coherence score of 0.969 for Amazon and identifying 368 topics. For Netflix, it achieved a coherence score of 0.99, identifying 402 topics, closely matching the outcomes of Meta Llama 3 8B. The Spotify dataset, however, saw a slightly lower coherence score of 0.962, with 356 topics identified,

[4] Such as Zero Shot on the Netflix dataset (Three-Label Scenario).

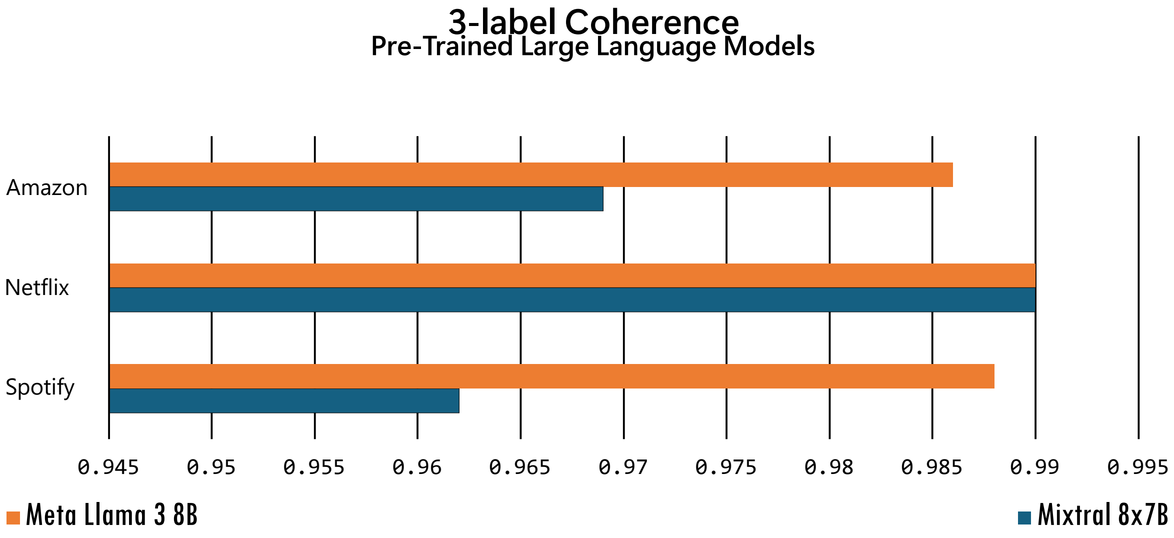


Figure 6.10: Bar Graph of Coherence Scores for All Datasets and Approaches in 3-Label Classification Using Meta Llama 3 8B and Mixtral 8x7B Models.

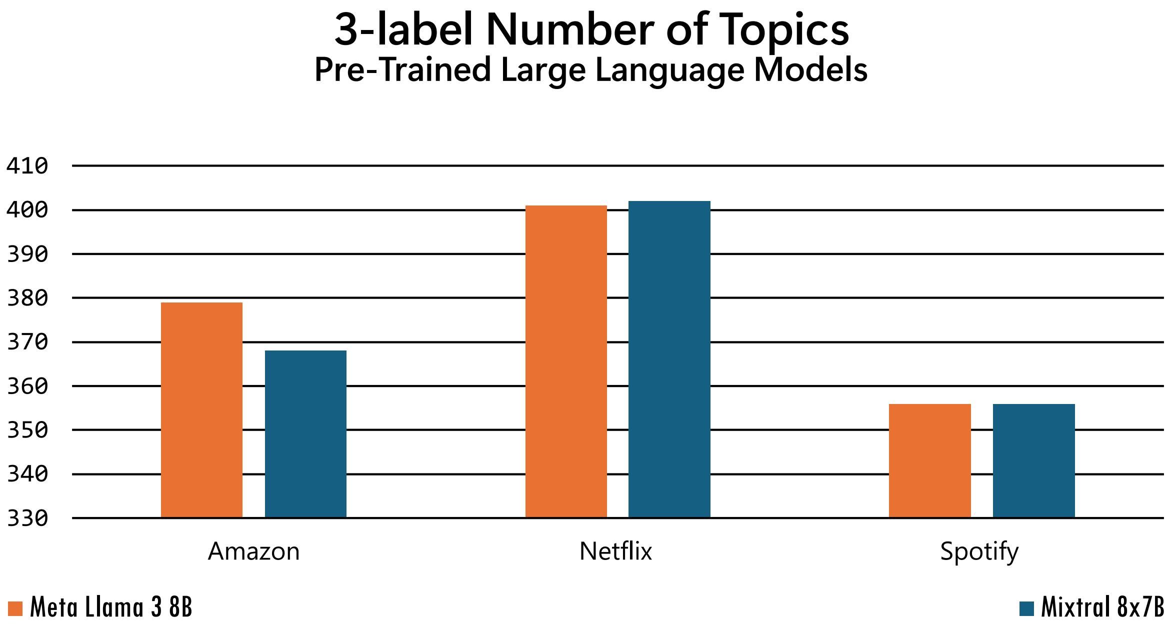


Figure 6.11: Bar Graph of the Number of Topics for All Datasets and Approaches in 3-Label Classification Using Meta Llama 3 8B and Mixtral 8x7B Models.

Dataset	Meta Llama 3 8B		Mixtral 8x7B	
	Coherence	Topics	Coherence	Topics
Amazon	0.988	379	0.969	368
Netflix	0.99	401	0.99	402
Spotify	0.986	356	0.962	356

Table 6.3: Topic Coherence and Number of Topics for Meta Llama 3 8B and Mixtral 8x7B Models in a 3-Label Classification Context.

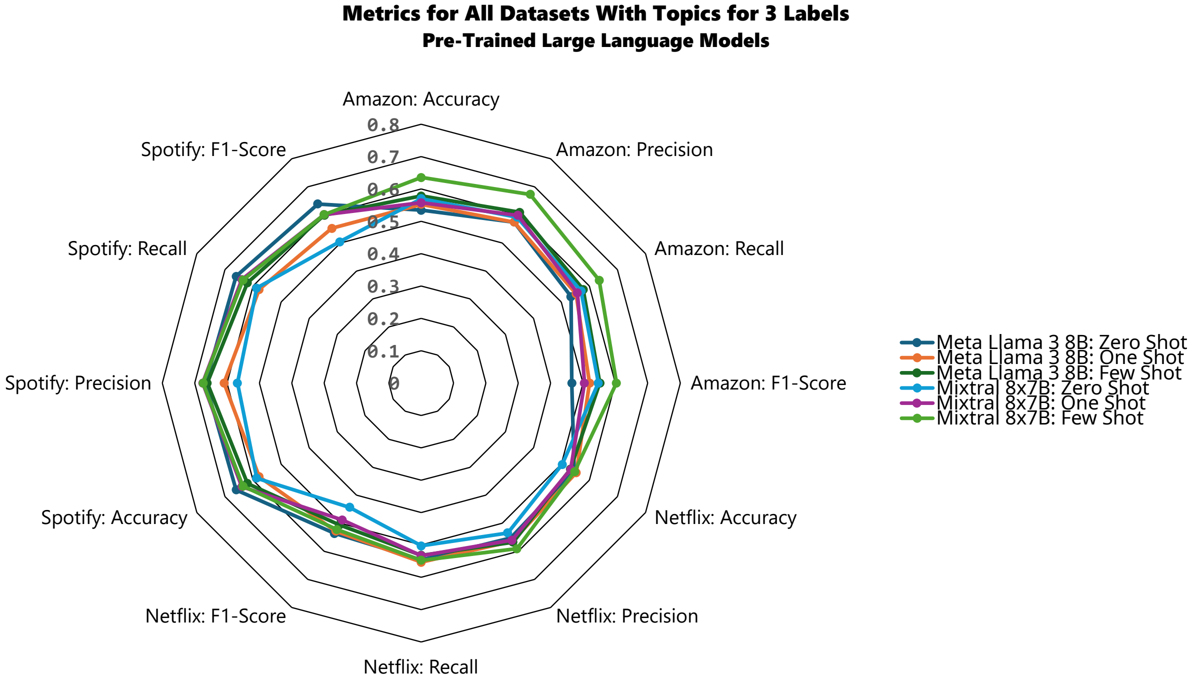


Figure 6.12: Radar Chart With Metrics for Meta Llama 3 8B and Mixtral 8x7B Models Across Datasets for 3-Label Sentiment Analysis With Topics.

reflecting a slight dip in the model's ability to maintain high coherence while generating a comparable number of topics to Meta Llama 3 8B.

In the three-label setting, the inclusion of topic modeling affected the results metrics compared to when no topics were used, as shown in Table 6.4 and Figure 6.12. The impact of topic modeling varied across different datasets and learning scenarios, with some benefiting more from integrating topics than others.

For the Spotify dataset, the Meta Llama 3 8B model achieved its highest F1 score in the Zero Shot scenario with a value of 0.639, followed by the Few Shot scenario at 0.599 and the One Shot scenario at 0.552. In the Amazon dataset, the model's outcome was highest in the Few Shot scenarios, achieving an F1 score of 0.552, with One Shot and Zero-Shot scenarios yielding slightly lower F1 scores of 0.52 and 0.534, respectively. Regarding the Netflix dataset, the model's F1 score peaked in the Zero Shot scenario at 0.536, with a slight decline observed in the One Shot scenario at 0.528 and a further reduction in the Few Shot scenario to 0.506. These results are reported in Table 6.4.

The Mixtral 8x7B model achieved its highest F1 score of 0.603 in the Few Shot learning scenario, followed by 0.547 in the Zero Shot scenario and 0.504 in the One Shot scenario,

Dataset	Metrics	Meta Llama 3 8B			Mixtral 8x7B		
		Zero Shot	One Shot	Few Shot	Zero Shot	One Shot	Few Shot
Amazon	Accuracy	0.534	0.552	0.578	0.57	0.557	0.635
	Precision	0.574	0.574	0.609	0.592	0.598	0.673
	Recall	0.534	0.552	0.578	0.57	0.557	0.635
	F1	0.465	0.52	0.552	0.547	0.504	0.603
Netflix	Accuracy	0.546	0.553	0.532	0.503	0.533	0.547
	Precision	0.553	0.564	0.567	0.534	0.56	0.591
	Recall	0.546	0.553	0.532	0.503	0.533	0.547
	F1	0.536	0.528	0.506	0.443	0.488	0.522
Spotify	Accuracy	0.66	0.578	0.62	0.587	0.64	0.636
	Precision	0.666	0.609	0.661	0.568	0.675	0.674
	Recall	0.66	0.578	0.62	0.587	0.64	0.636
	F1	0.639	0.552	0.599	0.504	0.601	0.601

Table 6.4: Evaluation of Meta Llama 3 8B and Mixtral 8x7B Models in 3-Label Topic-Informed Sentiment Analysis Across Datasets.

as observed in Table 6.4. For the Spotify dataset, the model recorded an F1 score of 0.601 in both the One Shot and Few Shot scenarios, with a lower F1 score of 0.504 in the Zero Shot scenario. In the case of the Netflix dataset, the model's F1 score was highest in the Few Shot scenario at 0.522, followed by 0.488 in One Shot and 0.433 in Zero Shot.

These results demonstrated the varying impact of topic modeling across different datasets and learning scenarios for both models. The Meta Llama 3 8B model showed particular strength in the Zero Shot scenario for the Spotify dataset. In contrast, the Mixtral 8x7B model excelled in Few Shot scenarios, especially for the Amazon dataset. The inconsistent results across datasets suggests that the effectiveness of topic modeling may be influenced by the unique characteristics of each dataset, such as content diversity, linguistic complexity, or sentiment distribution. This variability underscores the importance of careful model selection based on the specific requirements of each sentiment analysis task and dataset.

On average, Meta Llama 3 8B has an F1 score of 0.544. Specifically, in Zero Shot Learning, Meta Llama 3 8B has a mean score of 0.547; in One Shot Learning, it is 0.533; and in Few Shot Learning, it is 0.552. In contrast, Mixtral 8x7B has an average F1 score of 0.535. For Mixtral 8x7B, the scores are 0.498 in Zero Shot Learning, 0.531 in One Shot Learning, and 0.575 in Few Shot Learning. This shows that Meta Llama 3 8B surpasses Mixtral 8x7B, but not in all learning types.

The Wilcoxon signed-rank test was conducted to evaluate the null hypothesis: "Adding topics does not bring about a significant difference in F1 scores between Meta Llama 3 8B and Mixtral 8x7B for the three-label classification setting." The test returned a p-value of 1 (one). Since the p-value exceeds the significance threshold of 0.05, we cannot reject the null hypothesis, confirming no statistically significant difference in F1 scores between the two models.

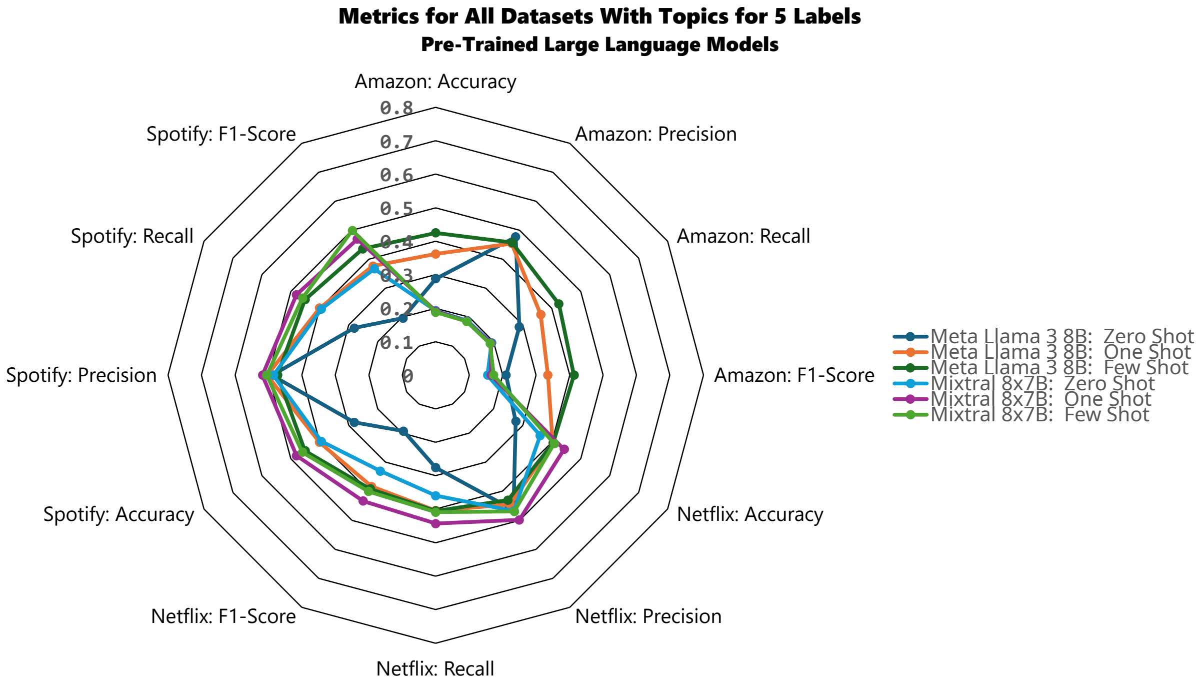


Figure 6.13: Radar Chart With Metrics for Meta Llama 3 8B and Mixtral 8x7B Models Across Datasets for 5-Label Sentiment Analysis With Topics.

Dataset	Meta Llama 3 8B		Mixtral 8x7B	
	Coherence	Topics	Coherence	Topics
Amazon	0.995	390	0.968	382
Netflix	0.996	398	0.965	384
Spotify	0.987	397	0.962	361

Table 6.5: Topic Coherence and Number of Topics for Meta Llama 3 8B and Mixtral 8x7B Models in 3-Label Classification.

Five-labels results

The complexity of sentiment classification increases significantly when the task is expanded to a five-class labeling scheme, where the sentiment spectrum is divided into very negative, negative, neutral, positive, and very positive classes. Figure 6.13 presents the obtained results.

The coherence scores for the five-label classification scheme (Table 6.5) highlight both the Meta Llama 3 8B and Mixtral 8x7B models' ability to maintain topic relevance while handling a more granular sentiment classification as demonstrated in Figure 6.14. Meta Llama 3 8B achieved coherence scores of 0.995 (Amazon), 0.996 (Netflix), and 0.987 (Spotify), identifying 390, 398, and 397 topics, respectively. Similarly, Mixtral 8x7B scored 0.968 (Amazon), 0.965 (Netflix), and 0.962 (Spotify), identifying 382, 384, and 361 topics, respectively. While Mixtral's coherence scores were slightly lower, both models demonstrated strong topic relevance, indicating their power in handling the increased complexity of the five-class scheme. The high coherence across datasets and the number of topics identified emphasize the models' capacity to maintain relevance even with more nuanced sentiment categories. The amount of topics are shown in Figure 6.15.

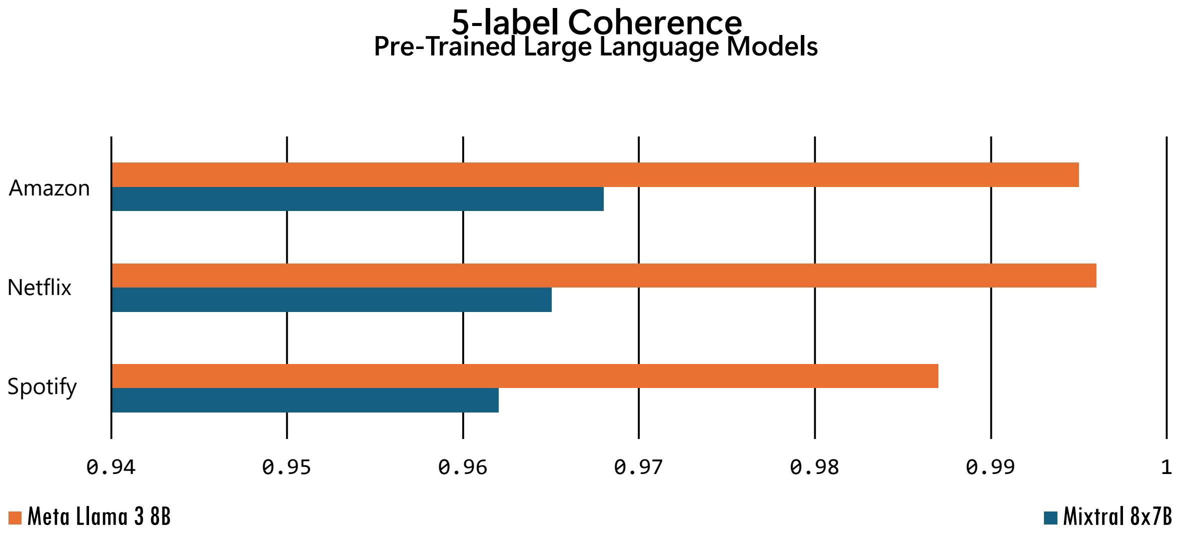


Figure 6.14: Bar Graph of Coherence Scores for All Datasets and Approaches in 5-Label Classification Using Meta Llama 3 8B and Mixtral 8x7B Models.

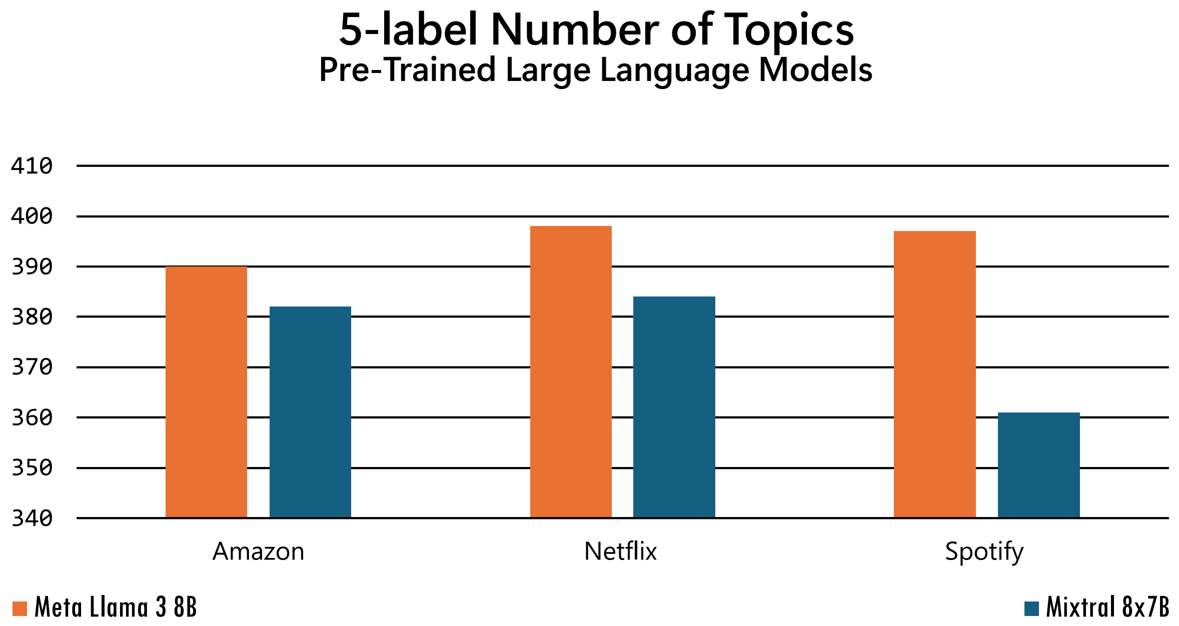


Figure 6.15: Bar Graph of the Number of Topics for All Datasets and Approaches in 5-Label Classification Using Meta Llama 3 8B and Mixtral 8x7B Models.

Dataset	Metrics	Meta Llama 3 8B			Mixtral 8x7B		
		Zero Shot	One Shot	Few Shot	Zero Shot	One Shot	Few Shot
Amazon	Accuracy	0.288	0.362	0.425	0.193	0.191	0.188
	Precision	0.477	0.454	0.457	0.188	0.186	0.185
	Recall	0.288	0.362	0.425	0.193	0.191	0.188
	F1	0.21	0.335	0.413	0.155	0.166	0.173
Netflix	Accuracy	0.276	0.404	0.405	0.36	0.443	0.409
	Precision	0.464	0.446	0.431	0.468	0.499	0.47
	Recall	0.276	0.404	0.405	0.36	0.443	0.409
	F1	0.193	0.384	0.392	0.331	0.434	0.401
Spotify	Accuracy	0.282	0.401	0.451	0.395	0.481	0.46
	Precision	0.49	0.504	0.473	0.485	0.518	0.505
	Recall	0.282	0.401	0.451	0.395	0.481	0.46
	F1	0.197	0.376	0.436	0.367	0.469	0.449

Table 6.6: Evaluation Metrics for Meta Llama 3 8B and Mixtral 8x7B Models in 5-Label Sentiment Analysis With Topics Across Datasets.

Coherence scores for three and five-label sentiment analysis tasks showed differences between Meta Llama 3 8B and Mixtral 8x7B models. In the three-label scenario, both maintained high coherence, with Meta Llama 3 8B slightly outperforming Mixtral 8x7B, especially for Spotify. Meta Llama 3 8B showed higher coherence across all datasets in the five-label scenario. At the same time, Mixtral 8x7B's scores were consistently lower, with the gap widening for Amazon and Netflix. While both models are compelling, Meta Llama 3 8B demonstrates superior topic coherence, particularly in more complex sentiment classification tasks.

The Meta Llama 3 8B model exhibits varied results in sentiment analysis tasks involving five labels, particularly when topic modeling is integrated.

In the Spotify dataset, Meta Llama demonstrates its robustness, achieving its highest F1 score of 0.436 in the Few Shot learning scenario, followed by scores of 0.376 in the One Shot scenario and 0.197 in the Zero Shot scenario. Similarly, the model maintains consistent outcomes in the Amazon dataset, attaining an F1 score of 0.413 in Few Shot, 0.335 in One Shot, and 0.21 in Zero Shot. However, the model's aftermath in the Netflix dataset is slightly lower, with its best F1 score of 0.392 occurring in the Few Shot scenario, followed by 0.384 in One Shot and 0.193 in Zero Shot.

These results suggest that while Meta Llama 3 8B is generally robust across different datasets, its influence can vary depending on the dataset and learning scenario. All these results are depicted in Table 6.6.

The Mixtral 8x7B model (Table 6.6)demonstrates notable strengths in sentiment analysis tasks, particularly within the Netflix and Spotify datasets when topic modeling is employed.

In the Spotify dataset, Mixtral 8x7B achieves its highest F1 score of 0.469 in the One Shot scenario, surpassing Meta Llama 3 8B in the same context. Additionally, Mixtral 8x7B attains a competitive F1 score of 0.449 in the Few Shot scenario, with a score of

0.367 in Zero-Shot, exceeding Meta Llama in this dataset. However, Mixtral 8x7B exhibits weaker execution on the Amazon dataset, with its highest F1 score reaching only 0.173 in the Few Shot scenario, significantly below Meta Llama's corresponding result, and scoring 0.166 in One Shot and 0.155 in Zero Shot. In the Netflix dataset, Mixtral 8x7B outmatch Meta Llama across all scenarios, achieving F1 scores of 0.434 in One Shot, 0.401 in Few Shot, and 0.331 in Zero Shot.

These findings suggest that while Mixtral 8x7B is highly effective in specific datasets and learning scenarios, mainly when fewer examples are provided, it demonstrates variability in its ability to generalize across different datasets.

Regarding averages, Meta Llama 3 8B has an average F1 score of 0.326. Specifically, in Zero Shot Learning, Meta Llama 3 8B has a mean score of 0.2; in One Shot Learning, it is 0.365; and in Few Shot Learning, it is 0.414. In contrast, Mixtral 8x7B has an average F1 score of 0.334. For Mixtral 8x7B, the scores are 0.284 in Zero Shot Learning, 0.356 in One Shot Learning, and 0.361 in Few Shot Learning. This shows that Mixtral 8x7B outruns Meta Llama 3 8B by a small margin, but not in all learning types.

The Wilcoxon signed-rank test was performed to determine whether adding topics affected F1 scores between Meta Llama 3 8B and Mixtral 8x7B in five-label classification. With a p-value of 0.82, this result shows no significant difference in F1 scores between the models.

6.3 Discussion

Due to the small sample size in this study, we emphasize descriptive statistics to compare Meta Llama 3 8B and Mixtral 8x7B across Zero, One, and Few Shot learning.

In the three-label classification without topics, Meta Llama 3 8B leads with an average F1 score of 0.59, especially excelling in Zero-Shot learning (0.6). Mixtral 8x7B scores lower overall at 0.521, though Few Shot learning (0.569) is its best. For five-label classification without topics, Mixtral 8x7B surpasses Meta Llama 3 8B, with an average F1 score of 0.36 compared to 0.27 for Meta Llama. Mixtral's Few Shot learning (0.372) also surpasses Meta Llama's best result.

For the three-label classification with topics, Meta Llama 3 8B achieves an average F1 score of 0.544, with Few Shot learning (0.552) being the best, while Mixtral 8x7B scores slightly lower at 0.535 but excels in Few Shot learning (0.575). For the five-label classification with topics, Meta Llama 3 8B has an average F1 score of 0.326, with Few Shot learning (0.414) being the highest. Mixtral 8x7B, with a similar average score of 0.334, significantly outperforms in Few Shot learning (0.361) compared to One Shot (0.356) and Zero Shot (0.284), showing superior outcomes in this more complex scenario.

When comparing Zero, One, and Few-shot learning methods, Few-shot learning consistently emerges as the most effective approach for both models across nearly all classification tasks. Meta Llama 3 8B excels in three-label classification without topics and behaves robustly with topics. Still, Mixtral 8x7B proved more effective for five-label tasks, especially when topics are included. Overall, Few Shot learning stands out as the most reliable learning method, and Mixtral 8x7B demonstrates more robust results in more complex tasks. At the same time, Meta Llama 3 8B excels in more straightforward settings.

Without topic modeling, the three-label scheme achieved an average F1 score of 0.555, while the five-label scheme scored 0.315. The Wilcoxon Signed-Rank Test produced an extremely small p-value of 7.63×10^{-6}, indicating a significant achievement difference favoring the three-label scheme. With topic modeling, the three-label scheme also stages better, with an average F1 score of 0.539 compared to 0.327 for the five-label scheme. The Wilcoxon Signed-Rank Test confirmed this result with a p-value of 7.63×10^{-6}, highlighting the consistent superiority of the three-label scheme.

The average F1 score without topics is 0.435, with 0.555 for the three-label scenario and 0.315 for the five-label scenario. This indicates that these models struggle more with complex sentiment classification without topic modeling. However, when topics are incorporated, the average F1 score improves to 0.429, showing enhanced outcomes, especially in the five-label scenario (0.327). The three-label average remains strong at 0.539, suggesting that topic modeling benefits transformer-based models in both classification scenarios.

The coherence values were quite high, approaching one, which is the optimal score. However, these high coherence values did not correlate with the sentiment analysis outcomes.

6.4 Final Remarks

Our findings underscored the potential of pre-trained large language models, like Meta Llama 3 8B and Mixtral 8x7B, in sentiment prediction – particularly when enhanced with topic modeling – and topic generation. These models demonstrated their potential across various learning configurations, with Few Shot learning consistently emerging as the most reliable approach for extracting nuanced sentiment information. However, the complexity introduced by the five-label classification scheme proved a significant challenge, especially in scenarios without topic modeling. In contrast, the simpler three-label scheme provided more consistent results across models, reinforcing its practicality for many sentiment classification tasks.

One of the key takeaways is the critical role topic modeling plays in enhancing the metrics of sentiment analysis models. Topic modeling improves sentiment classification, particularly in more complex tasks involving five sentiment labels. This reveals that both Meta Llama 3 8B and Mixtral 8x7B models benefited from integrating advanced topic modeling techniques, with improvements in F1 scores and overall classification results.

This study sets the stage for further exploration into the comparative strengths of small fine-tuned NLP models and pre-trained large language models in sentiment analysis. The upcoming Chapter delves deeper into these comparisons, offering a comprehensive discussion that integrates the results of the experiments with both small models (such as BERT and T5) and the larger models analyzed in this chapter. The next Chapter provides an integrated discussion, comparing the results from both small fine-tuned models and pre-trained large language models, and highlights the broader contributions of this research.

We highlight the study's limitations and suggest avenues for future investigations, guiding practitioners and researchers in selecting and optimizing models for sentiment analysis tasks in various contexts.

Chapter 7

Conclusion

This Chapter presents a conclusion to this M.Sc. Dissertation integrating a detailed discussion of the results obtained from sentiment analysis experiments with small fine-tuned NLP models – specifically BERT,T5 and BERTopic for Topic Modeling – and pre-trained large language models. We further discuss the practical implications of our findings and potential future work directions.

Our research evaluated the effectiveness of small fine-tuned NLP models, such as BERT and T5, alongside pre-trained large models, Meta Llama 3 8B and Mixtral 8x7B, for sentiment analysis under various scenarios, using both three-class and five-class sentiment classification scheme, with and without the integration of topic modeling techniques. Results were primarily assessed using F1 scores alongside other key metrics such as accuracy, precision, and recall.

Findings revealed BERT consistently outrun T5, showing consistently strong effectiveness, especially with three-label and both in scenarios with topic modeling and without, with the ATC and UCA being the most compelling topic modeling methods, which improved generalization and sentiment classification. Meta Llama 3 8B excelled in Zero-Shot learning tasks, while Mixtral 8x7B showed superior outcomes in Few-Shot learning, excelling particularly in complex five-class classification with topics.

Across most scenarios, the small fine-tuned models consistently outperformed the pre-trained large models, with BERT standing out in topic-informed and non-topic-based tasks. Our study found that the three-class sentiment classification scheme yielded better results than the five-class scheme. This was particularly true when topic modeling was not employed, where the more granular labeling in the five-class scheme often decreased classification metrics. However, the integration of topic modeling helped mitigate these challenges, especially in more complex tasks, enhancing model coherence and results in five-class classification. We provide further overall discussions in the following.

7.1 Overall Discussion

This study compared sentiment analysis models, exploring BERT, T5, Meta Llama 3 8B, and Mixtral 8x7B across different classification approaches and topic modeling techniques. Our findings highlighted the achievement of our investigation under these various configurations, emphasizing their contributions and limitations. BERT consistently exceeds T5 in three- and five-label classifications, with higher F1 scores across Amazon, Netflix, and Spotify datasets. Topic modeling techniques like UCA and ATC demonstrated superior results to LS and UTG, particularly in three-label classifications. Despite a slight decline when topics were incorporated, the three-label scheme remained more successful, yielding better overall F1 scores.

When comparing Meta Llama 3 8B and Mixtral 8x7B, Few-Shot learning consistently emerged as the most effective learning method. Meta Llama excelled in three-label classifications without topics, while Mixtral proved more effective in five-label classifications, mainly when topics were included. The results also indicated that topic modeling enhances model outcomes in more complex scenarios, with Mixtral exceeding Meta Llama in five-label settings. Overall, three-label schemes outperformed five-label configuration, and topic modeling benefited transformer models, especially in handling more detailed sentiment classifications.

The coherence values for the pre-trained large language models were significantly better than those for BERTopic, the small fine-tuned NLP technique. However, in both cases, it was not possible to determine how these coherence values affected sentiment analysis.

This section brings closure in answering this research's main question and addressing its main objective. Overall, the research demonstrates while that topic modeling can enhance sentiment analysis in complex five-label scenarios, especially with larger pre-trained language models (such as Mixtral), smaller fine-tuned models, particularly BERT, show superior results without topic.

The remainder of this section focuses on addressing the specific research questions and objectives.

Comparative Insights into Sentiment Analysis: Topic-Based vs. Non-Topic Approaches

When sentiment analysis results are compared with and without topics, the results vary significantly across different models and labeling schemes. Small fine-tuned NLP models (BERT and T5) tend to do better overall without topics, particularly in simpler sentiment classification tasks. Conversely, larger models (Meta Llama 3 8B and Mixtral 8x7B) improved topic modeling effectiveness, especially in more complex tasks. We found that the efficacy of topic-based sentiment analysis versus regular sentiment analysis depends on the specific model and task complexity. Specific research question RQ 1 and specific objective OB 1 were conveyed here.

Comparison of Small Fine-Tuned NLP Models vs. Pre-Trained Large Language Models in Sentiment Analysis

The Small Fine-Tuned NLP Models and Techniques approach achieves an overall average F1 score[1] of 0.555 when not considering topics. This achievement surpasses that of the

[1] With an average of 0.607 for three labels and 0.504 for five labels.

Pre-Trained Large Language Models approach, which has an overall average F1 score[2] of 0.435.

In the best topic-informed scenario (UCA), the Small Fine-Tuned NLP Models and Techniques approach shows an overall average F1 score[3] of 0.513. This is also better than the Pre-Trained Large Language Models approach, which has an overall average F1 score[4] of 0.433. These differences suggest that the small models provide slightly superior results in sentiment analysis tasks compared to the larger models. Here, specific research question RQ 2 and specific objective OB 2 were addressed.

Three-Label vs. Five-Label in Sentiment Classification and Topic Modeling Approaches Influence

The evaluation of small fine-tuned NLP models and pre-trained large language models across different sentiment classification tasks reveals notable differences when comparing three-label and five-label schemes.

In summary, across all evaluated configurations—regardless of the model type or topic modeling technique—the three-label scheme consistently demonstrated better results than the five-label scheme. The results, consistently supported by statistically significant p-values from the Wilcoxon Signed-Rank Tests, indicate that the three-label scheme is more effective for sentiment classification tasks. The segment regarding different sentiment classifications in specific research question RQ 3 and specific objective OB 3 was fully approached in this part.

UCA and ATC improved model generalization by consolidating topics across the corpus, showing superior results, this addresses the part regarding the effect of topic granularity in the specific research question RQ 3 and the specific objective OBJ 3.

7.2 Implications and Future Investigations

The findings of this M.Sc. dissertation carry significant implications for both academic research and practical applications in the field of NLP. For practitioners, the results clearly show that BERT remains a strong choice for sentiment classification tasks involving fewer sentiment labels and not requiring topic modeling. However, for more complex tasks–such as five-label classification–larger pre-trained large models like Meta Llama 3 8B and Mixtral 8x7B demonstrate their strengths, mainly when supported by Few Shot learning and topic modeling techniques like UCA and ATC.

Theoretically, the study deepens the understanding of how task complexity and dataset characteristics influence model metrics in sentiment analysis. It emphasized the relevance of careful model selection and configuration when dealing with nuanced sentiment classification tasks. The study also reveals the computational trade-offs of working with large pre-trained large models. The resource-intensive nature of these models, coupled with connection stability issues during the experimentation phase, underscores the need for

[2] The average for three labels is 0.555, while for five labels is 0.315.

[3] The average F1 score is 0.592 for three labels and 0.434 for five labels.

[4] For three labels, the average F1 score is 0.601, whereas it drops to 0.449 for five labels.

efficient deployment strategies in real-world applications. These insights are valuable for those working in sentiment analysis and for researchers developing NLP models for other tasks.

In addition to this research providing valuable insights into sentiment analysis, it opens up several promising avenues for future work. One direction is the exploration of additional models and datasets. Expanding the study to include other pre-trained large models like GPT-4 or RoBERTa or examining domain-specific datasets would help validate the generalizability of the findings across different languages, domains, and sentiment classification tasks.

Another potential area of exploration is advanced topic modeling techniques. Future studies could explore more sophisticated methods, exceptionally multilingual or multi-domain contexts. Multilingual datasets could enhance the research by allowing comparisons across languages, which may reveal different sentiment dynamics and topic relationships. This would provide richer insights and broaden the applicability of the findings. Integrating ABSA could further refine the understanding of sentiment related to specific aspects of products or services, making the analysis even more comprehensive. These avenues present exciting opportunities to deepen the understanding of sentiment analysis and its applications.

7.3 Limitations

Our research provided a detailed exploration of topic modeling and sentiment analysis using advanced models, but limitations need to be considered, particularly regarding the time and computational resources involved.

One of the main constraints was the time required to run the various models. While the experiments using BERTopic, BERT, and T5 were executed on the lab's[5] computational cluster, which provided sufficient resources for these models although time-consuming, the experiments with Meta Llama 3 8B and Mixtral proved far more challenging. These models were run using the Groq API, but the connection to the API was not consistently stable. Frequent interruptions and delays caused by connection issues significantly slowed down the process. Additionally, the API itself was sometimes slow, extending the duration of these experiments. This combination of technical challenges and the computationally intensive nature of large transformer models resulted in experiments taking longer than anticipated.

Given the limited time available to complete this M.Sc. project, these delays impacted the breadth of experiments predicted to be performed[6]. While it was possible to run several vital configurations and analyze the results, the time lost to technical difficulties meant that further exploration of additional models, datasets, or other settings was not possible within the scope of this study.

Furthermore, in some instances, the study's reliance on descriptive statistics was due

[5] Reasoning for Complex Data (Recod) laboratory. Recod.AI's website. Available at: <`https://recod.ai`>. Accessed on September 16, 2024.

[6] It was not possible to investigate the integration of topic modeling with BERTopic on Llama and Mixtral, nor the topics generated using Llama and Mixtral in BERT and T5.

to the relatively small sample size in parts of the topic-informed sentiment analysis. While the findings remain meaningful, a larger sample and more time would have allowed for a deeper, more comprehensive evaluation.

7.4 Contributions

This research provided several contributions to sentiment analysis. First, it offered a thorough empirical comparison between smaller fine-tuned models such as BERT and T5, and larger pre-trained large models like Meta Llama and Mixtral. The study illuminated their strengths and limitations across different sentiment classification tasks, offering clear evidence of their suitability for various scenarios.

By integrating topic modeling techniques, our research advanced the understanding of how topic modeling can enhance sentiment classification, especially in more granular five-label schemes with pre-trained large models. The detailed evaluation of Zero, One, and Few Shot learning configurations presented new insights into the optimal settings for pre-trained large models. Few-shot learning emerged as the most reliable configuration across most tasks.

Another contribution was the exploration of three-label versus five-label classification schemes, revealing the challenges of more granular sentiment analysis and providing evidence that simpler schemes yielded better outcomes. This research offered practical guidance for NLP practitioners and researchers, highlighting when and how to use different models, topic modeling techniques, and learning configurations for improved sentiment analysis.

Our conducted research led to the publication of a full article at an international conference:

- Araújo, Patrick; Haddadi, Seyed; Santos, Fillipe; Reis, Marcelo; Dos Reis, Julio Cesar. 2024. "Topic Modeling Influence in Sentiment Analysis from User-generated Product Reviews". In Proceedings of the IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Reggio Emilia, Italy, 2024.

7.5 Closure Remarks

This M.Sc. Dissertation advanced sentiment analysis by demonstrating how topic modeling techniques impact the proper classification. Findings revealed to which extent BERT consistently outrun T5, showing consistent results, especially with three-label configuration and scenarios with topic modeling and without. Meta Llama 3 8B excelled in Zero-Shot learning tasks, and Mixtral 8x7B showed superior outcomes in Few-Shot learning, excelling particularly in complex five-class classification with topics. We found that the small fine-tuned models consistently outperformed the pre-trained large models, with BERT standing out. Overall, all models performed better in the three classes. Our study provided practitioners with nuanced guidelines for selecting appropriate models and methods based on task complexity, highlighting the relevance of thoughtful model architecture and appropriate settings in achieving more precise and contextually relevant sentiment analysis.

Bibliography

[1] Nur Atiqah Sia Abdullah and Nur Ida Aniza Rusli. Multilingual Sentiment Analysis: A Systematic Literature Review. *Pertanika Journal of Science and Technology*, 29(1), 2021.

[2] Abeer Abuzayed and Hend Al-Khalifa. BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. *Procedia Computer Science*, 189:191–194, 2021.

[3] Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In Alexander Koller and Katrin Erk, editors, *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany, March 2013. Association for Computational Linguistics.

[4] Kefah Alissa and Omar Alzoubi. Financial sentiment analysis based on transformers and majority voting. In *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–4, Dec 2022.

[5] Amina Amara, Mohamed Ali Hadj Taieb, and Mohamed Ben Aouicha. Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. *Applied Intelligence*, 51(5):3052–3073, May 2021.

[6] Serpil Aslan. A deep learning-based sentiment analysis approach (mf-cnn-bilstm) and topic modeling of tweets related to the ukraine–russia conflict. *Applied Soft Computing*, 143:110404, 2023.

[7] Mohammed Attia, Younes Samih, Ali Elkahky, and Laura Kallmeyer. Multilingual Multi-class Sentiment Classification Using Convolutional Neural Networks. In *LREC 2018 Proceedings*, Miyazaki, Japan, May 2018. ELRA.

[8] Syed Muzamil Basha and Dharmendra Singh Rajput. Chapter 9 - survey on evaluating the performance of machine learning algorithms: Past contributions and future roadmap. In Arun Kumar Sangaiah, editor, *Deep Learning and Parallel Computing Environment for Bioengineering Systems*, pages 153–164. Academic Press, 2019.

[9] Manju Bhardwaj, Priya Mishra, Shikha Badhani, and Sunil K. Muttoo. Sentiment analysis and topic modeling of COVID-19 tweets of India. *International Journal of System Assurance Engineering and Management*, 15(5):1756–1776, May 2024.

[10] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, 2021.

[11] Jordan Boyd-Graber, Yuening Hu, and David Mimno. Applications of Topic Models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296, 2017.

[12] Yunna Cai, Fan Wang, Haowei Wang, and Qianwen Qian. Public sentiment analysis and topic modeling regarding ChatGPT in mental health on Reddit: Negative sentiments increase over time. *ArXiv*, 2023.

[13] Yong Chen, Hui Zhang, Rui Liu, Zhiwen Ye, and Jianying Lin. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, 163:1–13, January 2019.

[14] Biraj Dahal, Sathish A. P. Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9(1):24, December 2019.

[15] Tiago De Melo and Carlos M S Figueiredo. Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach. *JMIR Public Health and Surveillance*, 7(2):e24585, February 2021.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018.

[17] Qian Dong, Tingting Sun, Yan Xu, Xuguang Xu, Mei Zhong, and Kai Yan. Network public opinion sentiment analysis based on bert model. In *2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN)*, pages 662–666, Aug 2022.

[18] Roman Egger and Joanne Yu. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7:886498, May 2022.

[19] Paolo Fornacciari, Monica Mordonini, and Michele Tomaiuolo. A case-study for sentiment analysis on twitter. In *Workshop From Objects to Agents*, 2015.

[20] Muhammad Jauharul Fuadvy and Roliana Ibrahim. Multilingual Sentiment Analysis on Social Media Disaster Data. In *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, pages 269–272, Denpasar, Indonesia, October 2019. IEEE.

[21] Piyush Ghasiya and Koji Okamura. Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach. *IEEE Access*, 9:36645–36656, 2021.

[22] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv*, abs/2203.05794, 2022.

[23] Loni Hagen. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management*, 54(6):1292–1307, November 2018.

[24] Maoxin (Molson) Han. How does mobile device usage influence review helpfulness through consumer evaluation? Evidence from TripAdvisor. *Decision Support Systems*, 153:113682, February 2022.

[25] Tatsuya Ikeagami, Xin Kang, and Fuji Ren. Improvement of Japanese Text Emotion Analysis by Active Learning Using Transformers Language Model. In *2022 14th International Conference on Computer Research and Development (ICCRD)*, pages 171–177, Shenzhen, China, January 2022. IEEE.

[26] Hyeju Jang, Emily Rempel, David Roth, Giuseppe Carenini, and Naveed Zafar Janjua. Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis. *Journal of Medical Internet Research*, 23(2):e25431, February 2021.

[27] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, June 2019.

[28] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, 2024.

[29] Huang Jing and Cai Yang. Chinese text sentiment analysis based on transformer model. In *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*, pages 185–189, Jan 2022.

[30] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models.* Prentice Hall PTR, 3rd edition, 2024. Online manuscript released August 20, 2024.

[31] Suvarna Kadam and Vinay Vaidya. Review and analysis of zero, one and few shot learning approaches. In Ajith Abraham, Aswani Kumar Cherukuri, Patricia Melin, and Niketa Gandhi, editors, *Intelligent Systems Design and Applications*, pages 100–112, Cham, 2020. Springer International Publishing.

[32] Bradley Karas, Sue Qu, Yanji Xu, and Qian Zhu. Experiments with LDA and Top2Vec for embedded topic discovery on social media data—A case study of cystic fibrosis. *Frontiers in Artificial Intelligence*, 5:948313, August 2022.

[33] J Kazmaier and Jh Van Vuuren. Sentiment analysis of unstructured customer feedback for a retail bank. *ORiON*, 36(1), August 2020.

[34] Pooja Kherwa and Poonam Bansal. Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), 7 2019.

[35] Fotis Kitsios, Maria Kamariotou, Panagiotis Karanikolas, and Evangelos Grigoroudis. Digital Marketing Platforms and Customer Satisfaction: Identifying eWOM Using Big Data and Text Mining. *Applied Sciences*, 11(17):8032, August 2021.

[36] Niharika Prasanna Kumar and Srivathsan Mohan. Sentiment analysis of russo-ukrainian war using twitter text corpus. In *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, volume 1, pages 1–5, April 2023.

[37] V. Kumar, Divya Ramachandran, and Binay Kumar. Influence of new-age technologies on marketing: A research agenda. *Journal of Business Research*, 125:864–877, March 2021.

[38] Josip Kunsabo and Jasminka Dobša. A systematic literature review on topic modelling and sentiment analysis. In *Central European Conference on Information and Intelligent Systems*, pages 371–380. Faculty of Organization and Informatics Varazdin, 09 2022.

[39] Shanghao Li, Zerong Xie, Dickson K. W. Chiu, and Kevin K. W. Ho. Sentiment analysis and topic modeling regarding online classes on the reddit platform: Educators versus learners. *Applied Sciences*, 13(4), 2023.

[40] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A Critical Review of Recurrent Neural Networks for Sequence Learning, 2015.

[41] Bing Liu. *Sentiment Analysis and Opinion Mining.* Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham, 2012.

[42] Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48(4):499–527, December 2017.

[43] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[44] Lindung Parningotan Manik, Harry Susianto, Arawinda Dinakaramani, Niken Pramanik, and Totok Suhardijanto. Can lexicon-based sentiment analysis boost performances of transformer-based models? In *2023 7th International Conference on New Media Studies (CONMEDIA)*, pages 314–319, Dec 2023.

[45] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[46] Banafsheh Mehri, Martin Trépanier, and Yves Goussard. Multilingual text classification on social media data for incident alert in subway transportation network. *Centre interuniversitaire de recherche sur les reseaux d'entreprise, la logistique et le transport (CIRRELT).*

[47] Pooja Mehta and Dr.Sharnil Pandya. A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific & Technology Research*, 9:601–609, 2020.

[48] Yida Mu, Peizhen Bai, Kalina Bontcheva, and Xingyi Song. Addressing Topic Granularity and Hallucination in Large Language Models for Topic Modelling, 2024.

[49] R. Nagaraj, C.R Rohith Adithya, Sakalabathula Sri Chakra Teja, and Deepika T. Identifying the influences behind the linkedin posts using topic modeling and sentiment analysis. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*, pages 1–7, 2024.

[50] Ramesh Narwal and Himanshu Aggarwal. Predicting Online Game-Addicted Behaviour with Sentiment Analysis Using Twitter Data. In Anuradha Tomar, Hasmat Malik, Pramod Kumar, and Atif Iqbal, editors, *Machine Learning, Advances in Computing, Renewable Energy and Communication*, volume 768, pages 505–517. Springer Singapore, Singapore, 2022.

[51] Ogobuchi Daniel Okey, Ekikere Umoren Udo, Renata Lopes Rosa, Demostenes Zegarra Rodríguez, and João Henrique Kleinschmidt. Investigating chatgpt and cybersecurity: A perspective on topic modeling and sentiment analysis. *Computers Security*, 135:103476, 2023.

[52] Eunil Park, Yeonju Jang, Jina Kim, Nam Jeong Jeong, Kunwoo Bae, and Angel P. Del Pobil. Determinants of customer satisfaction with airline services: An analysis of customer feedback big data. *Journal of Retailing and Consumer Services*, 51:186–190, November 2019.

[53] Abhilash Pathak, Sudhanshu Kumar, Partha Pratim Roy, and Byung-Gyu Kim. Aspect-based sentiment analysis in hindi language by ensembling pre-trained mbert models. *Electronics*, 10(21), 2021.

[54] Keval Pipalia, Rahul Bhadja, and Madhu Shukla. Comparative analysis of different transformer based architectures used in sentiment analysis. In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 411–415, 2020.

[55] Kundeti Naga Prasanthi, Rallabandi Eswari Madhavi, Degala Naga Sai Sabarinadh, and Battula Sravani. A novel approach for sentiment analysis on social media using bert roberta transformer-based models. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–6, April 2023.

[56] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1427–1445, March 2022.

[57] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of machine learning research*, 2019.

[58] Ankita Rane and Anand Kumar. Sentiment classification system of twitter data for us airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 01, pages 769–773, 2018.

[59] Asutosh Rath, B. Hridaya, Duggaraju Vimala, and Jossy George. Multilingual sentiment analysis of youtube live stream using machine translation and transformer in nlp. In *2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)*, pages 1–5, Oct 2022.

[60] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[61] Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*, 2014.

[62] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020.

[63] Siri Amanda Rääf, Johanna Knöös, Fisnik Dalipi, and Zenun Kastrati. Investigating learning experience of moocs learners using topic modeling and sentiment analysis. In *2021 19th International Conference on Information Technology Based Higher Education and Training (ITHET)*, pages 01–07, 2021.

[64] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA, February 2015. Association for Computing Machinery.

[65] Kazuhiro Seki, Masahiko Shibamoto, and Takashi Kamihigashi. Topic-sentiment analysis of central bank press conferences: Boj case study. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2861–2865, 2023.

[66] Mamoona Shabbir and Muhammad Majid. Sentiment analysis from urdu language-based text using deep learning techniques. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–5, Feb 2024.

[67] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K. Reddy. Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 1105–1114, Lyon, France, 2018. ACM Press.

[68] Carson Sievert and Kenneth Shirley. LDAvis: A method for visualizing and interpreting topics. In Jason Chuang, Spence Green, Marti Hearst, Jeffrey Heer, and Philipp Koehn, editors, *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[69] Sunju Park, Seung-Yong Lee, and Seungwha Chung. The Effects of Online Product Reviews on Sales Performance: Focusing on Number, Extremity, and Length. *Journal of Distribution Science*, 17(5):85–94, May 2019.

[70] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307, 06 2011.

[71] Soonh Taj, Baby Bakhtawer Shaikh, and Areej Fatemah Meghji. Sentiment analysis of news articles: A lexicon based approach. In *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–5, Jan 2019.

[72] Amira Samy Talaat. Sentiment analysis classification system using hybrid BERT models. *Journal of Big Data*, 10(1):110, June 2023.

[73] Joseph J Thompson, Betty Hm Leung, Mark R Blair, and Maite Taboada. Sentiment analysis of player chat messaging in the video game StarCraft 2: Extending a lexicon-based model. *Knowledge-Based Systems*, 137:149–162, December 2017.

[74] Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. AnchiBERT: A Pre-Trained Model for Ancient Chinese Language Understanding and Generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Shenzhen, China, July 2021. IEEE.

[75] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023.

[76] Danny Valdez, Andrew C. Pickett, and Patricia Goodson. Topic Modeling: Latent Semantic Analysis for the Social Sciences: Topic Modeling. *Social Science Quarterly*, 99(5):1665–1679, November 2018.

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *CoRR*, 2017.

[78] Aarshitha Vemulapalli and Anudeep Peddi. A comparative study of twitfeel and transformer-based techniques for the analysis of text data for sentiment classification. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, volume 6, pages 683–688, Sep. 2023.

[79] Markos Viggiato, Dayi Lin, Abram Hindle, and Cor-Paul Bezemer. What Causes Wrong Sentiment Classifications of Game Reviews? *IEEE Transactions on Games*, 14(3):350–363, September 2022.

[80] Jason Wei and Kai Zou. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, 2019.

[81] Wikipedia contributors. Cross entropy — Wikipedia, the free encyclopedia, 2023. [Online; accessed 30-March-2023].

[82] Wikipedia contributors. Language model — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Language_model&oldid=1146119319`, 2023. [Online; accessed 31-March-2023].

[83] Wikipedia contributors. Erik erikson — Wikipedia, the free encyclopedia, 2024. [Online; accessed 24-November-2024].

[84] Wikipedia contributors. K-means clustering — Wikipedia, the free encyclopedia, 2024. [Online; accessed 27-October-2024].

[85] Wikipédia. Aprendizado de máquina — wikipédia, a enciclopédia livre, 2022. [Online; accessed 18-outubro-2022].

[86] Wikipédia. Aprendizagem profunda — wikipédia, a enciclopédia livre, 2023. [Online; accessed 12-fevereiro-2023].

[87] Wikipédia. Grandes modelos de linguagem — wikipédia, a enciclopédia livre, 2024. [Online; accessed 17-abril-2024].

[88] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[89] Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July 2020. Association for Computational Linguistics.

[90] Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A Survey on Neural Topic Models: Methods, Applications, and Challenges. *Artificial Intelligence Review*, 2024.

[91] Jianfei Yu, Kai Chen, and Rui Xia. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):1966–1978, July 2023.

[92] David Zendle, Rachel Meyer, and Nick Ballou. The changing face of desktop video game monetisation: An exploration of exposure to loot boxes, pay to win, and cosmetic microtransactions in the most-played Steam games of 2010-2019. *PLOS ONE*, 15(5):e0232780, May 2020.

Appendix A

Appendix: Article Generated In This MS.c. Context

Araújo, Patrick; Haddadi, Seyed; Santos, Fillipe; Reis, Marcelo; Cesar Dos Reis, Julio, “Topic Modeling Influence in Sentiment Analysis from User-generated Product Reviews,” 2024 IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Reggio Emilia, Italy, 2024.