NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA

PROJECT REPORT

on

INTEL PRODUCTS SENTIMENT ANALYSIS FROM ONLINE REVIEWS

Submitted by:

Harshith Kolluru	1NT22CS073
Gagan S Kunkanad	1NT22CS067

Under the Guidance of

Dr. Sudhir Shenai

Associate Professors, Dept. of ISE, NMIT

Department of Computer Science and Engineering

(Accredited by NBA Tier-1)

2023-2024

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM

Department of Computer Science and Engineering

(Accredited by NBA Tier-1)

CERTIFICATE

This is to certify that the Project Report on **"Intel Products Sentiment Analysis from Online Reviews"** is an authentic work carried out by **Harshith Kolluru (1NT22CS073) and Gagan S Kunkanad (1NT22CS067)** Bonafede students of Nitte Meenakshi Institute of Technology, Bangalore in partial fulfillment for the award of the degree of Bachelor of Engineering in Computer Science and Engineering of Visvesvaraya Technological University, Belagavi during the academic year 2023-2024**.** It is certified that all corrections and suggestions indicated during the assessment have been incorporated in the report.

Dr. Sudhir Shenai

Associate Professor, Dept. ISE,

NMIT Bangalore

Abstract

This project focuses on sentiment analysis of Intel products based on user and expert reviews gathered from various online sources spanning the last 3-5 years. Leveraging the VADER lexicon for sentiment scoring and employing extensive text preprocessing techniques, the analysis categorizes sentiments into positive, negative, competition-related, and future expectation categories. Additionally, dimensionality reduction using Singular Value Decomposition (SVD) enhances feature extraction from textual data. The project culminates in the application of a Random Forest classification model to predict sentiment labels, providing insights into user sentiments and recommendations for product improvement based on analysis outcomes.

The first step in the process involves the collection and aggregation of user and expert reviews from multiple online platforms. This data, which encompasses opinions on various Intel products, is meticulously cleaned and preprocessed to ensure consistency and relevance. Text preprocessing techniques such as tokenization, stop-word removal, stemming, and lemmatization are employed to standardize the text and remove noise. This comprehensive preprocessing stage is critical for enhancing the accuracy of the sentiment analysis, as it ensures that the input data is both clean and uniform.

Following preprocessing, the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon is applied to assign sentiment scores to each review. VADER is particularly well-suited for analyzing sentiments expressed in social media and other online content due to its ability to capture both the intensity and the polarity of sentiments. Reviews are categorized into four distinct sentiment classes: positive, negative, competition-related, and future expectations. This categorization enables a nuanced understanding of the sentiments associated with Intel products, highlighting areas of strength as well as aspects that may require improvement.

To further enhance the analysis, Singular Value Decomposition (SVD) is utilized for dimensionality reduction. SVD helps in extracting the most informative features from the text data by reducing its dimensionality, thus making the subsequent analysis more efficient and focused. The refined features are then fed into a Random Forest classification model, which is trained to predict sentiment labels based on the processed text data. The Random Forest model, known for its robustness and accuracy, provides valuable insights into user sentiments, revealing patterns and trends that can inform Intel's product development and marketing strategies. The project ultimately offers actionable recommendations for Intel, derived from a thorough and sophisticated analysis of user and expert reviews.

Acknowledgement

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. I express my sincere gratitude to our Principal **Dr. H. C. Nagaraj**, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD**, Dr. S Meenakshi Sundaram.** for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better project.

I would like to extend my deepest gratitude to the individuals and institutions that supported me throughout this project. Firstly, I am immensely grateful to the **Intel Unnati Industrial Training Program** for providing the resources and platform necessary for this project. Their commitment to fostering innovation and technical skills among students has been instrumental in the successful completion of this work.

I would like to express my heartfelt thanks to my mentor **Dr. Sudhir Shenai** for their unwavering guidance, insightful feedback, and constant encouragement. Their expertise and mentorship have been invaluable in navigating the complexities of this project.

I also wish to thank my college lecturer for their support and for facilitating this project as part of the curriculum. Their dedication to nurturing student potential has been a significant driving force behind my academic achievements.

Finally, I extend my gratitude to my family and friends for their continuous support and encouragement. Their belief in my abilities has motivated me to strive for excellence in every endeavor.

Table of Contents

Chapter 1: Introduction

Sentiment analysis has become a crucial tool in understanding user opinions and feedback. With the exponential growth of online reviews and social media, businesses and researchers have access to vast amounts of data that can provide insights into customer satisfaction and preferences. This project aims to harness the power of sentiment analysis to analyze user reviews and derive meaningful conclusions.

In the contemporary landscape of consumer feedback, understanding sentiment and extracting meaningful insights from product reviews are pivotal for businesses striving to enhance customer satisfaction and product competitiveness. Sentiment analysis, in conjunction with topic modeling, provides a robust framework to distill sentiments, identify prevalent topics, and gauge consumer expectations and concerns.

This report delves into the application of advanced natural language processing (NLP) techniques to analyze sentiments and derive topics from a dataset of product reviews. The study employs state-of-the-art tools such as NLTK's Vader for sentiment analysis and sklearn's Latent Dirichlet Allocation (LDA) for topic modeling. By categorizing reviews based on sentiment polarity and identifying key discussion topics, this research aims to provide actionable insights for product improvement and market strategy formulation.

The primary objective of this project is to perform sentiment analysis on user reviews, categorizing the sentiments as positive, negative, or neutral. By doing so, we aim to provide businesses with valuable insights that can help improve their products or services. Understanding customer sentiment is essential for making informed business decisions and enhancing customer experience.

In this chapter, we introduce the project's scope, objectives, and significance. We discuss the importance of sentiment analysis in today's data-driven world and outline the structure of the report. This chapter sets the stage for the detailed discussion of methodologies, results, and conclusions that follow.

Chapter 2: Literature Review

2.1 Sentiment Analysis: A Journey from Simple to Sophisticated

The field of sentiment analysis has witnessed a fascinating evolution, driven by the explosion of digital text data and the relentless march of Natural Language Processing (NLP) advancements. Early forays into sentiment analysis often relied on rudimentary keyword-based methods. These methods involved the creation of lists of positive and negative words, with the sentiment of a text snippet being determined by the presence or absence of these words. However, this simplistic approach proved ineffective in capturing the nuances of human language. Sarcasm, for instance, would often be misinterpreted as positive sentiment due to the presence of positive words used ironically. Additionally, these methods struggled to account for the impact of context and sentiment modifiers on the overall polarity of a text.

The limitations of keyword-based methods paved the way for more sophisticated approaches leveraging the power of machine learning and deep learning.

- ### 2.1.1 Lexicon-Based Methods and The Rise of VADER:

Lexicon-based sentiment analysis methods address some of the shortcomings of keyword-based approaches. These methods utilize sentiment lexicons, which are comprehensive databases of words with pre-assigned sentiment scores. A lexicon might assign a score of +1 to a word like "happy" and -1 to a word like "sad." Sentiment analysis tools then analyze the text, assigning a sentiment score based on the sum of the scores of the individual words it contains.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a prominent example of a lexicon-based sentiment analysis model. VADER goes beyond simple word scores by incorporating grammatical and lexical heuristics. For instance, VADER recognizes that the sentiment of the word "not" reverses the sentiment of the preceding word. Additionally, VADER accounts for capitalization and punctuation, recognizing that exclamation points often intensify sentiment. VADER's ability to handle these nuances makes it particularly effective for analyzing informal text prevalent on social media and online review platforms.

- **2.1.2 Deep Learning Techniques: Unveiling Complexities in Text Data:**

The field of sentiment analysis has witnessed a revolution with the advent of deep learning techniques. Deep learning models, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), are capable of learning complex patterns within textual data. Unlike lexicon-based methods, deep learning models do not rely on pre-defined features or sentiment lexicons. Instead, they are trained on massive amounts of labeled text data, where each text snippet is assigned a sentiment label (e.g., positive, negative, neutral). Through this training process, the deep learning model learns to identify subtle patterns in word usage, sentence structure, and context, enabling it to achieve superior sentiment analysis accuracy.

One of the strengths of RNNs is their ability to handle sequential data, making them well-suited for analyzing text where the order of words can significantly impact sentiment. For instance, the sentence "The movie was great, despite the bad acting" conveys a positive sentiment despite containing the negative word "bad." RNNs can learn these intricacies of sentence structure and context, leading to more accurate sentiment analysis.

CNNs, on the other hand, excel at identifying local patterns within text data. This makes them adept at capturing sentiment expressed through emojis, hashtags, and other stylistic elements commonly used in social media communication.

2.2 Topic Modeling: Unveiling Hidden Themes within the Textual Labyrinth

Topic modeling has emerged as a powerful tool for uncovering latent themes within vast collections of textual data (corpora). Imagine a large corpus of customer reviews – topic modeling helps us identify underlying topics discussed within these reviews. For instance, a topic might emerge around positive customer experiences with product features, while another topic might focus on negative sentiment related to customer service interactions.

Latent Dirichlet Allocation (LDA) is a widely adopted topic modeling technique. LDA operates under the assumption that documents within a corpus are composed of a mixture of underlying topics, each characterized by a probability distribution over words. By analyzing word co-occurrence patterns within the corpus, LDA identifies these latent topics and estimates the topic distribution for each document. Essentially, LDA helps us categorize documents based on

the thematic content they share.

2.3 The Power of Fusion: Sentiment Analysis and Topic Modeling

Recent research emphasizes the significant benefits of integrating sentiment analysis with topic modeling. This combined approach offers a more granular level of insight when analyzing textual data. Imagine analyzing a collection of social media posts about a new product launch. Sentiment analysis alone might reveal that a significant portion of the conversation holds negative sentiment. However, by integrating topic modeling, we can delve deeper and understand the specific aspects of the product launch that are generating negative sentiment. For instance, topic modeling might reveal that negative sentiment clusters around the product's pricing strategy, while positive sentiment focuses on innovative features.

Chapter 3: Problem Solved and Objectives

The primary problem addressed by this project is the need for businesses to understand customer sentiment from large volumes of user reviews. Traditional methods of analyzing customer feedback, such as surveys and focus groups, are often limited in scope and can be time-consuming and expensive. Sentiment analysis offers a scalable and efficient alternative, enabling businesses to gain insights from vast amounts of text data.

Our project aims to solve this problem by developing a system for automated sentiment analysis of user reviews. The specific objectives are to collect a large dataset of user reviews, preprocess the data to ensure accuracy, apply sentiment analysis to categorize the sentiments, and visualize the results to provide actionable insights. By achieving these objectives, we aim to demonstrate the value of sentiment analysis in enhancing customer understanding and decision-making.

In addition to sentiment analysis, we aim to perform topic modeling to identify key themes in the user reviews. This will help businesses understand the specific aspects of their products or services that are most frequently mentioned by customers. By combining sentiment analysis with topic modeling, we provide a comprehensive solution for analyzing user feedback.

The primary objective of this study is to perform sentiment analysis and topic modeling on a corpus of product reviews to achieve the following:

1. **Sentiment Analysis**: Classify reviews into positive, negative, and neutral sentiments using NLTK's Vader sentiment analyzer.
2. **Topic Modeling**: Employ LDA to identify prevalent topics and themes within the reviews.
3. **Insight Generation**: Extract actionable insights regarding product performance, customer satisfaction, and areas for improvement based on sentiment and topic analysis.

By addressing these objectives, this research endeavors to equip businesses with data-driven insights that facilitate informed decision-making and enhance consumer-centric strategies.

Chapter 4: Detailed Description Of Work

This chapter provides a detailed description of the work undertaken for sentiment analysis and topic modeling of user reviews.

4.1 Data Acquisition

This section details the process of collecting the user review data:

- **4.1.1 Data Sources:** Specify the online platforms targeted for web scraping. Mention specific platforms like Amazon, Yelp, and TripAdvisor used to gather reviews.
- **4.1.2 Data Description:** Describe the collected data in detail. This includes:
    - Type of reviews (e.g., product reviews)
    - Timeframe for data collection
    - Volume of data obtained (number of reviews)
    - Any relevant metadata collected alongside the review text (e.g., ratings)

4.2 Data Preprocessing

This section explains the steps taken to clean and prepare the text data for analysis:

- **4.2.1 Text Cleaning:** Explain the importance of data cleaning and the techniques used to remove noise and inconsistencies. This might involve:
    - Removing stop words (common words with minimal meaning)
    - Removing punctuation and special characters
- **4.2.2 Text Normalization:** Describe the text normalization techniques used for consistency:
    - Converting all text to lowercase
    - Applying stemming or lemmatization (reducing words to their root forms) - optional, depending on the analysis requirements
- **4.2.3 Tokenization:** Briefly explain the process of breaking down the text into individual words (tokens) for further analysis.

4.3 Sentiment Analysis

This section details the sentiment analysis process:

- **4.3.1 VADER Model Selection:** Justify the choice of the VADER model. Explain its suitability for analyzing informal text prevalent in user reviews. Briefly describe how VADER assigns sentiment scores based on a sentiment lexicon and grammatical heuristics.
- **4.3.2 Sentiment Categorization:** Explain how the sentiment scores from VADER were used to categorize user reviews. This might involve:
    - Describing the specific sentiment score ranges used for classifying reviews as positive, negative, or neutral.

4.4 Topic Modeling with LDA

This section explains the topic modeling process using Latent Dirichlet Allocation (LDA):

- **4.4.1 Latent Dirichlet Allocation (LDA):** Introduce the concept of LDA and its role in topic modeling. Briefly explain how LDA works by identifying latent topics within the text data based on word co-occurrence patterns.
- **4.4.2 Determining the Number of Topics:** Describe the process used to determine the optimal number of topics for your analysis. This might involve iterative modeling and coherence analysis techniques.

4.5 Categorization and Insight Generation

This section details the process of extracting insights from the data:

- **4.5.1 Thematic Grouping:** Explain how reviews were categorized into thematic groups following sentiment analysis and topic modeling. Examples of thematic groups could be product performance, customer service, and user expectations.
- **4.5.2 Correlating Sentiment and Topics:** Describe the approach used to correlate sentiment trends with topics. This step is crucial for extracting insights into consumer sentiment and preferences based on the identified themes within the reviews.

Chapter 5: System Architecture

This chapter delves into the system architecture designed for our sentiment analysis and topic modelling pipeline. The architecture comprises several interconnected components, each playing a vital role in transforming raw user reviews into actionable insights. By adopting a modular approach, this system facilitates the efficient and scalable analysis of large textual datasets.

5.1 Data Acquisition: Gathering Reviews from the Web

The initial stage involves collecting user reviews from relevant online platforms. Web scraping techniques automate this process, efficiently extracting valuable data. Here, we leverage Python libraries like BeautifulSoup and Scrapy:

- **BeautifulSoup:** This library excels at parsing HTML and XML content retrieved from websites. It provides intuitive functions for navigating the structure of web pages, allowing us to target specific elements containing review text, ratings, and other pertinent metadata.
- **Scrapy:** Scrapy offers a robust framework for building scalable web crawlers. It streamlines the data extraction process by handling website navigation, data parsing, and storage efficiently. Particularly for complex websites or large-scale data collection, Scrapy provides a robust and maintainable solution.

Data Storage: The collected reviews, ratings, and metadata are then stored in a structured format, typically a relational database (e.g., MySQL, PostgreSQL) or a NoSQL database (e.g., MongoDB) depending on project requirements. This structured storage facilitates efficient data access and manipulation during subsequent processing stages.

5.2 Data Preprocessing: Preparing Text for Analysis

The raw review text obtained from web scraping might contain noise and inconsistencies that can hinder the accuracy of sentiment analysis and topic modeling. The preprocessing stage addresses these issues, transforming the text data into a suitable format for analysis.

- **Noise Removal:** This involves eliminating irrelevant elements from the text, such as stop words (common words like "the", "a", "an") that convey little sentiment or thematic meaning. Libraries like NLTK (Natural Language Toolkit) provide pre-built stop word lists in various languages.
- **Text Normalization:** This step ensures consistency in the text data. Common normalization techniques include:
    - Lowercasing all text: This eliminates the influence of case sensitivity on sentiment analysis and topic modeling algorithms.
    - Stemming or Lemmatization (optional): These techniques reduce words to their root forms, improving the ability of the algorithms to identify underlying themes and sentiment. Stemming is a simpler approach, while lemmatization aims for more accurate morphological analysis. The choice between stemming and lemmatization depends on the specific requirements of the analysis.
- **Tokenization:** The preprocessed text is then segmented into individual words (tokens) for further analysis. Tokenization allows sentiment analysis models and topic modeling algorithms to work with the text data in a structured manner.

5.3 Sentiment Analysis: Unveiling Sentiment through VADER

This stage focuses on classifying the sentiment expressed within user reviews. We leverage the VADER (Valence Aware Dictionary and sEntiment Reasoner) model, a lexicon-based sentiment analysis tool particularly adept at handling informal text commonly found in social media and online reviews.

VADER assigns sentiment scores to text based on a pre-defined sentiment lexicon and grammatical heuristics. The lexicon contains a list of words with associated sentiment polarities (positive, negative, or neutral). Additionally, VADER considers grammatical elements like negation ("not") and capitalization to refine sentiment analysis accuracy.

The sentiment analysis component outputs sentiment scores for each review. These scores can then be used to categorize reviews as positive, negative, or neutral based on predefined sentiment score thresholds.

5.4 Topic Modeling: Unveiling Latent Themes

Topic modeling is a powerful technique for uncovering hidden thematic structures within large collections of text data. In our system architecture, Latent Dirichlet Allocation (LDA) serves as the primary topic modeling algorithm.

Latent Dirichlet Allocation (LDA):

LDA operates under the assumption that documents within a corpus (collection of text data) are composed of a mixture of underlying topics. Each topic is characterized by a probability distribution over the words that might appear within it. By analyzing word co-occurrence patterns across the corpus, LDA identifies these latent topics and estimates the topic distribution for each document. Essentially, LDA helps us categorize documents based on the thematic content they share.

For instance, consider a corpus of customer reviews for a smartphone. LDA might identify topics such as "camera quality," "battery life," and "display performance." Each review would then be assigned a probability distribution over these topics, indicating the prevalence of each theme within the specific review.

Determining the Number of Topics (K):

A crucial step in topic modeling involves selecting the optimal number of topics (K) to be identified within the data. Here are some approaches to guide this selection:

- Domain Knowledge: Leveraging your understanding of the domain and the expected themes within the reviews can provide a good starting point for selecting K.
- Perplexity: Perplexity is a metric that measures how well a topic model fits the unseen data. Lower perplexity values generally indicate a better fit. We can iterate through different K values, evaluating perplexity for each model. A "knee" in the perplexity curve often suggests the optimal number of topics, where the perplexity starts to increase rapidly as we add more topics that may not be capturing significant thematic distinctions.
- Coherence Scores: Various coherence scores can be used to assess the semantic quality

of the topics identified by the model. Higher coherence scores indicate that the words within a topic are semantically related and form a meaningful theme. Common coherence scores include Coherence Value (C_V), Uniformity (Umass), and Normalized Pointwise Mutual Information (NPMI). We can evaluate these scores for different K values to identify the model that generates the most coherent topics.

5.5 Categorization and Insight Generation

Following sentiment analysis and topic modeling, we can derive valuable insights from the processed data.

- Thematic Grouping: Reviews can be categorized into thematic groups based on the dominant topics identified through LDA. This allows us to analyze sentiment trends within specific themes, providing a more granular understanding of customer opinions.
- Correlating Sentiment and Topics: By correlating sentiment scores with prevalent topics, we can gain deeper insights into customer sentiment and preferences. For example, we might identify a correlation between negative sentiment and the topic of "battery life." This suggests that customers who mentioned battery life in their reviews expressed more negative sentiment, potentially highlighting an area for improvement in the product.

These insights can be instrumental for businesses in various ways:

- Product Development: Identifying recurring themes and sentiment trends can inform product development efforts by highlighting areas where the product excels or falls short of customer expectations.
- Customer Service: Understanding the root causes of customer dissatisfaction through thematic analysis can guide customer service strategies to better address customer concerns.
- Marketing and Sales: Insights from sentiment analysis and topic modeling can be used to tailor marketing campaigns and sales strategies to resonate with customer preferences and address any pain points identified within the reviews.

Chapter 6: Methodology

This chapter outlines the methodological approach adopted for sentiment analysis and topic modeling of user reviews. The methodology can be broadly categorized into three stages: Data Collection and Preprocessing, Sentiment Analysis, and Topic Modeling.

6.1 Data Collection and Preprocessing

This stage involves gathering the user review data and preparing it for analysis.

- **6.1.1 Data Acquisition:**
    - Web scraping techniques are employed to collect user reviews from various online platforms (e.g., Amazon, Yelp, TripAdvisor).
    - The collected data includes review text, ratings, and other relevant metadata.
- **6.1.2 Data Preprocessing:**
    - Text cleaning techniques are applied to remove noise and inconsistencies from the text data. This includes:
        - Removing stop words (common words with minimal meaning)
        - Removing punctuation and special characters
    - Text normalization is performed to ensure consistency:
        - Converting all text to lowercase
        - Applying stemming or lemmatization (optional, reduces words to their root forms)
    - Tokenization is used to break down the text into individual words (tokens) for further analysis.

6.2 Sentiment Analysis

This stage focuses on classifying the sentiment expressed within user reviews.

- **6.2.1 Sentiment Analysis Model Selection:**
    - The VADER (Valence Aware Dictionary and sEntiment Reasoner) model is

chosen due to its suitability for analyzing informal text prevalent in user reviews.
    - VADER assigns sentiment scores to text based on a sentiment lexicon and grammatical heuristics.
- **6.2.2 Sentiment Categorization:**
    - The sentiment scores obtained from VADER are used to categorize user reviews into distinct classes (e.g., positive, negative, neutral).
    - Specific sentiment score ranges are defined to classify reviews into their respective categories.

6.3 Topic Modeling

This stage involves identifying latent themes within the user reviews using Latent Dirichlet Allocation (LDA).

- **6.3.1 Text Vectorization:**
    - Text data is converted into numerical vectors using a technique like TF-IDF (Term Frequency-Inverse Document Frequency).
    - TF-IDF weighting helps account for the importance of words based on their frequency within a document and rarity across the entire corpus.
- **6.3.2 LDA Modeling:**
    - Latent Dirichlet Allocation (LDA) is applied to the TF-IDF matrix to identify latent topics within the corpus.
    - LDA analyzes word co-occurrence patterns to group words into thematically related topics.
- **6.3.3 Topic Interpretation:**
    - The identified topics are interpreted based on the distribution of words within each topic.
    - Coherence scores are employed to evaluate the semantic quality of the topics.

Chapter 7: Skills and Knowledge Gained

This chapter details the professional and systemic skill sets acquired throughout this data science project, with a particular focus on the domain of Natural Language Processing (NLP).

7.1 Proficiency in Python Programming for Data Science Workflows

The project necessitated the development of proficiency in Python programming, a cornerstone skill for data science professionals. Our focus centred on mastering essential libraries that streamline various stages of the data science workflow:

- **Web Scraping with BeautifulSoup and Scrapy:** We harnessed the capabilities of BeautifulSoup and Scrapy libraries to automate data acquisition from websites. These libraries facilitated the efficient collection of user reviews, eliminating the need for manual data extraction and significantly improving data collection scalability.
- **Data Manipulation and Analysis with Pandas:** Pandas emerged as a vital tool for data manipulation and analysis. By leveraging Pandas' functionalities, we were able to transform the collected review data into a structured format (e.g., DataFrame). This structured organisation significantly enhanced the efficiency of subsequent data processing and analysis stages.
- **Natural Language Processing with NLTK:** The Natural Language Toolkit (NLTK) library provided a comprehensive suite of tools for NLP tasks. We utilised NLTK to perform essential text pre-processing steps, including the removal of stop words (common words with minimal semantic meaning) and the application of stemming or lemmatization techniques. These pre-processing steps are crucial for ensuring that NLP algorithms can effectively understand the underlying meaning within textual data.

7.2 Deepening Understanding of Natural Language Processing Techniques

The project provided a valuable platform to delve deeper into the intricacies of NLP techniques:

- **Sentiment Analysis: Unveiling Emotional Tone in Text:** We explored sentiment analysis, a technique that enables the programmatic classification of emotional sentiment

within textual data. Through this exploration, we gained a comprehensive understanding of how VADER assigns sentiment scores based on a pre-defined sentiment lexicon and grammatical heuristics.

- **Topic Modelling: Discovering Latent Themes:** We investigated topic modelling, a powerful NLP technique that facilitates the identification of latent themes within large textual datasets. Latent Dirichlet Allocation (LDA) served as the primary topic modelling algorithm used in this project. LDA analyses word co-occurrence patterns to group words into thematically related topics. This technique provides valuable insights into the overarching themes and topics discussed within user reviews.

7.3 Strengthening Data Analysis and Visualization Skills

The project fostered the development of robust data analysis and visualisation skills:

- **Data Analysis Expertise:** We honed our ability to interpret the results of sentiment analysis (sentiment scores) and topic modeling (topic distributions). This involved meticulous analysis of patterns and trends within the data to extract meaningful insights regarding customer sentiment and the key themes emerging from the user reviews.
- **Effective Data Visualization:** We acquired proficiency in creating clear and informative data visualizations using Matplotlib. These visualizations, encompassing elements like bar charts or word clouds, served as effective communication tools, enabling us to present the project's findings to a broader audience in a readily comprehensible manner.

By systematically applying these acquired skills and knowledge sets, we were able to successfully navigate the various stages of the data science project, ultimately transforming raw user reviews into actionable insights.

Chapter 8: Results

This chapter presents the key findings derived from the sentiment analysis and topic modeling stages of the project.

8.1 Sentiment Analysis Results

The sentiment analysis classified the user reviews into three categories: positive, negative, and neutral. Here's a breakdown of the distribution:

- **Positive Sentiment: 82%** of the reviews expressed positive sentiment towards the product. These reviews highlighted the product's strengths, favorable aspects, and overall satisfaction with factors like CPU performance, temperature control, and compatibility.
- **Negative Sentiment: 9%** of the reviews conveyed negative sentiment. These reviews pinpointed areas requiring improvement or addressed specific customer concerns, such as encountering issues with motherboards, power consumption, or encountering BSOD (Blue Screen of Death) errors.
- **Neutral Sentiment: 7%** of the reviews exhibited neutral sentiment. These reviews offered a balanced perspective without expressing strong emotional polarity towards the product, possibly focusing on providing general information or requesting compatibility details.

8.2 Topic Modeling Results

Topic modeling successfully identified **four** distinct topics or themes within the corpus of product reviews. These topics represent the overarching conversations and areas of focus within the reviews.

The four identified topics include:

- **Topic A: CPU Performance and thermals (i9, performance, temperature, cooler):** This topic centered around discussions regarding the performance of the central processing unit (CPU), particularly focusing on processors like the i9. Users discussed aspects like processing power, gaming performance, and thermal efficiency (heat

generation and cooling). Reviews within this topic might mention benchmarks, gaming experiences, and the effectiveness of cooling solutions (fans, liquid coolers) for managing thermals.

- **Topic B: Motherboard Compatibility and Power Consumption (motherboard, case, power, issues):** This topic focused on discussions about motherboard compatibility and overall system power consumption. Users might have explored questions about whether the motherboard supports specific CPU models (e.g., ASUS motherboards and Intel CPUs), explored potential issues with power requirements or wattage limitations, or inquired about troubleshooting steps related to encountered compatibility problems.
- **Topic C: CPU Generation and Support (Intel, gen, 14th, 13th, 12th, support, errors):** This topic pertained to discussions about CPU generation and compatibility. Users might have expressed interest in knowing if the product supports the latest generations of Intel CPUs (e.g., 14th gen), inquired about compatibility with specific CPU generations, or mentioned encountering errors related to CPU support.
- **Topic D: PC Building Troubleshooting (PC, build, games, BSOD, error, FPS):** This topic centered around troubleshooting common issues encountered during PC building, particularly those related to the CPU. Users might have shared experiences with encountering Blue Screen of Death (BSOD) errors, discussed troubleshooting steps for resolving performance issues like low FPS (frames per second) in games, or sought solutions for resolving build problems potentially related to CPU compatibility.

The topics identified through topic modeling were validated using coherence scores. These scores ensure the relevance and coherence of the topics, indicating that the topics accurately reflect the underlying thematic structures within the reviews.

By analyzing the sentiment distribution and the identified topics, we gain valuable insights into customer perception of the product. Positive sentiment indicates overall satisfaction with CPU performance and thermals, while negative sentiment highlights concerns regarding compatibility or troubleshooting issues. Topic modeling provides a deeper understanding of the specific aspects of the CPU and related components that are generating customer interest and discussion. These insights can be used to inform product development efforts, improve customer service by addressing common compatibility questions and troubleshooting steps, and tailor marketing campaigns to effectively target customer needs and concerns.

Chapter 9: Conclusion

This project successfully integrated sentiment analysis and topic modeling to extract valuable insights from user reviews. The positive sentiment (82%) indicated overall customer satisfaction with the product, particularly regarding CPU performance, thermals, and compatibility. Topic modeling revealed four key themes: CPU performance and thermals, motherboard compatibility, CPU generation support, and PC building troubleshooting.

These findings demonstrate the power of NLP in transforming user reviews into actionable intelligence. Businesses can leverage this intelligence to optimize product strategies by focusing on areas highlighted in positive reviews and addressing concerns raised in negative ones. Sentiment analysis can also guide customer service efforts to proactively address potential issues and enhance customer satisfaction. Understanding customer preferences gleaned from topic modeling allows businesses to adapt their offerings and marketing to stay ahead of the competition.

In today's data-driven economy, customer feedback is a cornerstone for business growth and innovation. As NLP methodologies advance and AI-powered analytics become more prevalent, sentiment analysis and topic modeling will play an even greater role in shaping strategic business initiatives and fostering consumer-centric practices. By embracing these data-driven approaches, businesses can effectively translate the voice of the customer into actionable insights for sustained success.

REFERENCES

- **Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment analysis: Measuring opinions. Procedia Computer Science, 45, 808-814. https://doi.org/10.1016/j.procs.2015.03.159**
- **Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89. https://doi.org/10.1145/2436256.2436274**
- **Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016**
- **Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135. https://doi.org/10.1561/1500000011**
- **Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46. https://doi.org/10.1016/j.knosys.2015.06.015**
- **Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., & Valencia-García, R. (2017). Sentiment analysis on tweets about diabetes: An aspect-level approach. Computational and Mathematical Methods in Medicine, 2017, 5140631. https://doi.org/10.1155/2017/5140631**
- **Singla, J., & Verma, A. (2019). Sentiment analysis on product reviews. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 568-573). IEEE. https://doi.org/10.1109/CONFLUENCE.2019.8776927**
- **Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: A survey. International Journal of Advanced Research in Computer Science and Software Engineering, 2(6), 282-292.**
- **Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253. https://doi.org/10.1002/widm.1253**