Trends Analysis on Azure Cloud Reviews

Project

By

Jolly Ogbolè

July 2023

Introduction: Trends Analysis on Azure Cloud Services

In today's digital landscape, cloud computing has revolutionized the way businesses operate, offering scalable and flexible solutions for various computing needs. Among the leading cloud service providers, Microsoft Azure stands out as a comprehensive platform that empowers organizations with a vast array of services, enabling them to leverage the power of the cloud and drive innovation and business growth. As more and more businesses adopt Microsoft Azure and migrate their applications to the cloud, there is an increasing demand to understand the trends and patterns of its cloud services. Extracting valuable insights from the vast amounts of unstructured data, such as customer feedback, product reviews, and social media interactions, is crucial for Azure's continuous improvement and enhancing customer satisfaction.

Text mining, a subfield of natural language processing (NLP), plays a pivotal role in analyzing and processing unstructured textual data to uncover significant trends and patterns. By employing NLP techniques, we can gain deeper insights into preferences, and sentiments expressed by Azure users, thereby aiding Microsoft's efforts in refining their cloud services to better cater to their customers' needs and drive revenue growth.

Problem Statement and Research

In this project, I attempt to explore the world of text mining and NLP in the context of Microsoft Azure cloud feedback. The objective is to analyze the textual data available from various sources related to Azure cloud services and derive valuable information to understand customer experiences, identify emerging trends.

Background Research and Research Question:

Research question: What are the key trends of Azure Cloud Services?

Azure Cloud Services generated $22.1 billion in revenue in the third quarter of fiscal year 2023, representing 16% of Microsoft's total revenue. Analyzing trends and insights on Azure using Natural Language Processing (NLP) techniques provides valuable insights for decision-making. Previous studies have leveraged NLP techniques in relevant domains. (Kim et al. 2020) analyzed over 55,000 airline reviews using Latent Dirichlet Allocation (LDA) model and identified 27 dimensions of satisfactions. Also, (Sutherland, K Kiatkawsin 2020) analyzes text data of 104,161 online reviews of Korean accommodation customers by using Latent Dirichlet Allocation (LDA), to find which topics of interest guests find important.

By analyzing user sentiments and feedback, identifying areas for improvement, and exploring emerging opportunities, the organization can understand Azure's strengths, address weaknesses, and stay ahead of the competition. Leveraging NLP allows for a comprehensive evaluation of Azure Cloud Services, enabling informed strategies and improved user experiences.

I sequence the implementation of this project in the following order:

1. **Data Collection:** Gather diverse textual data from online forums, social media platforms, customer reviews, blogs, and other relevant sources.
2. **Preprocessing and Text Cleaning:** Use NLP techniques to preprocess the raw data by tokenizing, removing stop words, stemming, and lemmatizing, making it suitable for analysis.
3. **Topic Modeling**: Implement algorithms to uncover latent themes and topics discussed in the data, providing insights into popular features and functionalities of Microsoft Azure.
4. **Sentiment Analysis:** Analyze sentiments over time to understand how user opinions change in response to product updates and new features.

5. **Visualization:**Report findings through visualizations and discussions to effectively communicate complex information and facilitate stakeholders' understanding.

The Dataset

Capterra is a website that helps businesses find and evaluate software and business services. It has over 900 software categories and over 1 million reviews from real users. Businesses can use Capterra to compare different software solutions, read reviews, and find the best fit for their needs. The dataset we use for this project is an aggregation of over 1000 real user reviews, scrapped and aggregated from this high impact source, capterra.com.

Steps For Cleaning the Data

To clean the data and create a normalized text corpus, I implement the following procedures:

- Tokenization: Split the text into individual words or tokens. This step helps in analyzing the text at a more granular level.
- Lowercasing: Convert all the tokens to lowercase. This ensures that words in different cases are treated as the same, eliminating any inconsistencies due to capitalization.
- Removing Punctuation: Strip out any punctuation marks from the text. Punctuation usually doesn't carry much meaning in text mining tasks and can be safely removed.
- Stop Word Removal: Remove common words that do not contribute much to the overall meaning of the text, such as articles (e.g., "a," "an," "the"), prepositions, and conjunctions. This step reduces the noise in the data.
- Removing Numbers: Remove numeric characters from the text, as they might not be relevant for many text mining tasks. However, there might be cases where numbers carry important information, in which we opt to keep the numbers..

- Handling Special Characters: Handle special characters, symbols, or non-alphanumeric characters based on the context and goals of the text mining project. For example, you might remove or replace emoticons, special symbols, or HTML tags if present.
- Handling Abbreviations and Contractions: Expand abbreviations and contractions to their full forms. This helps in maintaining consistency and improves the accuracy of subsequent analyses.
- Lemmatization or Stemming: Reduce words to their base or root forms. This step helps in standardizing the words and reducing vocabulary size.
- Removing Irrelevant Words: Remove any domain-specific or irrelevant words that don't contribute to the analysis. These words can include product-specific jargon, brand names, or any other terms that are not of interest.
- Handling Spelling Errors: Correct any obvious spelling mistakes using techniques like spell-checking or using a pre-built dictionary. This step ensures consistency and improves the accuracy of subsequent analyses.

Details of the Methodology

The proposed methodology is the Latent Dirichlet Allocation (LDA) topic modeling. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a three-layer Bayesian probability model, which can identify the semantic topic information in large-scale document sets or corpus. LDA is an unsupervised machine learning technique, and it has recently emerged as the method of choice for working with large collections of text documents. Model selection criteria takes into consideration the domain area, dataset complexity and reliable model performance metrics. The initial step involved tokenization, converting the text into

tokens. Subsequently, all text was converted to lowercase, and further steps encompassed the removal of stop words, punctuation marks, and tokens with a length shorter than three characters or longer than twenty-five characters from the text corpus. In the final step, text mining functions were applied to facilitate knowledge analysis and visualization.

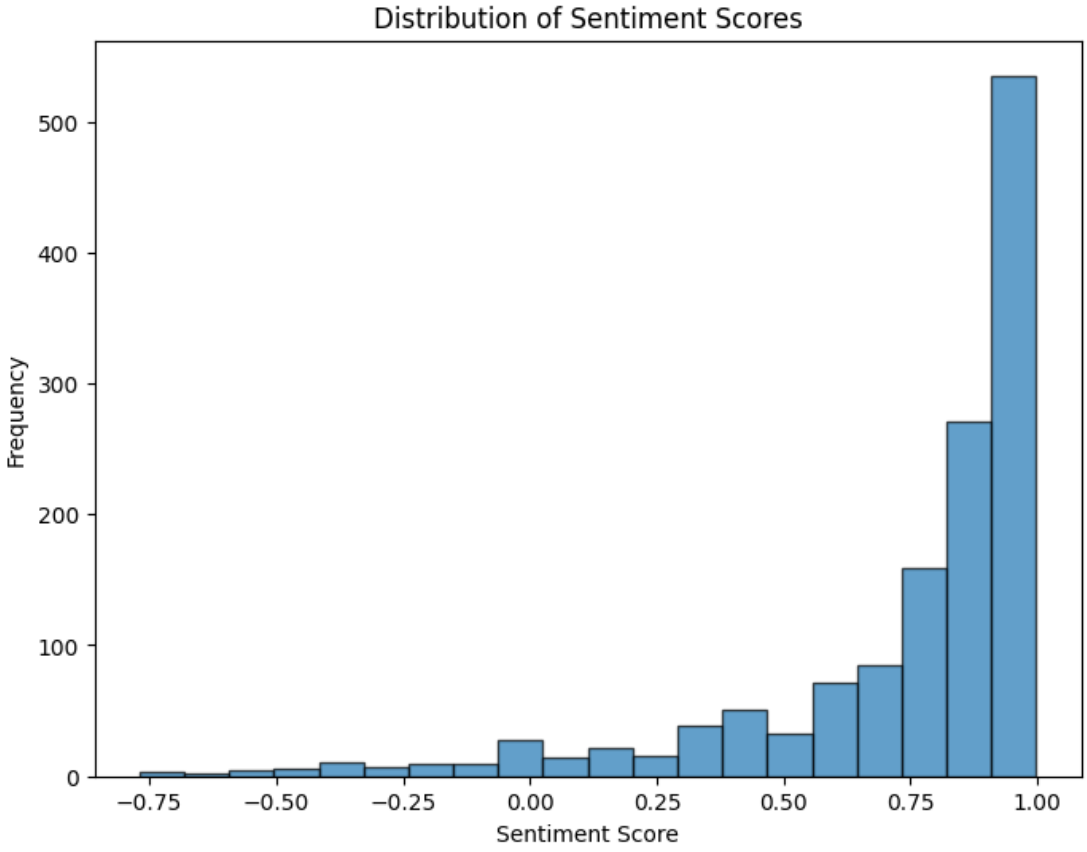
Implementation

Our customized text mining procedure enables the extraction of valuable information from unstructured data, facilitating knowledge analysis and visualization. Through text pre-processing, data cleaning, and feature extraction, we transform and apply structure to the data. Having fitted the structure to the tokenized corpus, I proceeded to deploy the LDA model on the corpus, and then visualize the initial output. After the initial findings, it was necessary to fine-tune the corpus and the parameters of the model to improve the accuracy and validity of the results.

Results

Sentiment Analysis

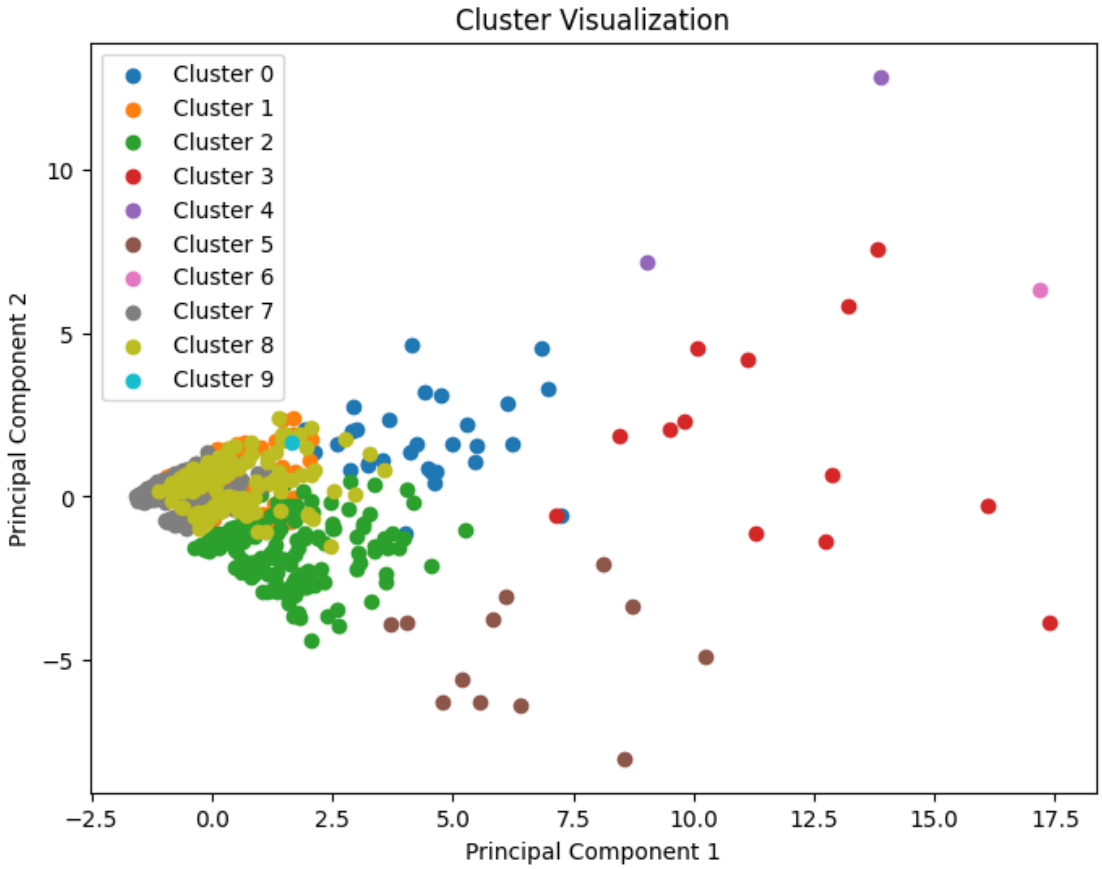
For the Sentiment Analysis I used python's SentimentIntensityAnalyzer from python's natural language toolkit library. By defining the function sentiment_scores, I obtained a prediction of sentiment scores for each review that fall on a sentiment score scale of -1 to 1. Then, after compiling the sentiment_scores for each review of our dataset, I visualize using matplotlib. This technique was very useful and insightful on determining the distribution of reviews sentiments. I report sentiment distribution below.



The histogram above is left skewed and shows the majority of the sentiment score above 0.75, This suggests an overall positive experience from the users of Azure.

Clustering Analysis

I used k-means clustering to identify document clusters within the data. The rationale for utilizing this technique is that we can identify meaningful and useful qualities in our dataset by finding documents that are more similar to each other than documents in other clusters. Where no document can be a member of more than one cluster. For my data, I decided to use 10 clusters, hence k = 10. I apply principal Component Analysis (PCA) for dimensionality reduction and utilize matplotlib to visualize a scatter plot of the clusters.


Cluster Visualization
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Principal Component 2
Principal Component 1
10
5
0
−5
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

LDA Model

The LDA model is implemented in tandem with the clustering analysis to allow for cluster labeling assignments based on the topics identified. I began by analyzing the documents in each of the 10 clusters, which outputted the dominant topic for each cluster.

Our meaningful labels for each of our 10 topics are: Application Development and Data Management, Comparing Free Cloud Solutions, Cost and Resource Management, Web Application Features and Performance, Server Support and Feature Usage, Business Security and Cloud Services, Security and Pricing Selection, Server and Application Performance in the Workplace, Service Support, and User Experience and Ease of Use.

Below I report the topical labels and the words they are composed of.

Table 1:

Topic	1. Application Development and Data management	2. Comparing Free Cloud Solutions	3. Cost and Resource Management	4. Web Application Features and Performance	5. Server Support and Feature Performance
Word	Product	Service	Time	Easy	Good
	Data	Using	Need	Application	Time
	Application	Free	Cost	Service	Server
	Machine	Solution	Data	Web	Really
	Virtual	AWS	Storage	Great	Used
	Management	Little	Resource	Feature	Support
	Development	Like	Feature	Issue	Lot
	Easy	Product	Company	Like	Would
	Integration	Bit	Server	Fast	Feature
	Work	Google	Pay	App	Like

\

Topic	6. Business Security and Cloud Services	7. Security and Pricing Selection	8. Server and Application Performance in the Workplace	9. Service Support	10. User Experience and Ease of Use
Word	Service	Service	Service	Easy	Service
	Business	Need	Server	Service	Like
	Feature	Cost	Virtual	Feature	Make
	User	Like	Machine	Support	Using
	Computing	Feature	Work	Like	Experience
	Security	High	Company	User	Tool
	Offer	Pricing	Cost	Need	User
	Range	Available	Time	Lot	Easy
	Application	Product	Great	Great	Much
	Provides	Security	Application	Good	Easier

Heatmap:


```
Top 5 words for Topic 0: Application Development: service, feature, need, application, user
Top 5 words for Topic 1: Cloud Solutions: easy, product, like, thing, great
Top 5 words for Topic 2: Cost and Resource Management: easy, user, cost, good, like
Top 5 words for Topic 3: Web Application Performance: service, like, application, data, using
Top 5 words for Topic 4: Server Support and Features: server, virtual, service, easy, machine
```

In the heatmap above, we observe a dominance of topic 0: Application Development in cluster 1 and cluster 3. Yet across all five clusters, topic 0: application development appears to be noticeably represented relative to other topic labels. The cell values represent the degree to which specific documents are present in the given cluster of documents.

Summary

In this research, I utilized a mix of text mining techniques to effectively uncover key trends. By leveraging LDA Model and Clustering analysis, I gained insights into specific tropical areas and elicited thematic understanding of the product reviews. These topic labels and their topic words offer insights for future Azure innovation direction, product and features iterations and improvements. And consequently creating opportunities for growth of the Microsoft Azure Cloud product.