Data Science and Engineering (DSE) Record, Volume 6, issue 1.

Sentiment Analysis and Data Visualization for Customer Satisfaction Survey in Healthcare Services

Natchayar Saosuwan [1] and Karn Patanukhom [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
natchayar_s@cmu.ac.th

Abstract. The healthcare sector is becoming more competitive, requiring businesses to understand consumer needs through sentiment analysis of feedback. This study analyzed feedback from Sriphat Medical Center to assess satisfaction (satisfied/dissatisfied) across eight aspects, including service process, staff behavior, and medical expertise. Using Natural Language Processing (NLP) and machine learning with Bag-of-Words and the Term Frequency-Inverse Document Frequency (TF-IDF) techniques, the best-performing model was a linear SVM with 95.8% accuracy in satisfaction classification and 77.4% in aspect classification.

Keywords: Natural Language Processing, machine learning, Bag-of-Words, the Term Frequency-Inverse Document Frequency

1 Introduction

Modern marketing is completely different from how it used to be. In the past, marketing primarily focused on production, as there were few manufacturers and demand exceeded supply. This resulted in minimal competition between companies. Later, as production increased and the market expanded, economic growth accelerated, leading businesses to pay more attention to and emphasize marketing.

As a result, the marketing concept shifted—from focusing solely on production and the product itself—to prioritizing customer satisfaction and needs, as well as addressing societal concerns [1].

The healthcare business sector is now facing intense competition due to the increasing number of private healthcare facilities, giving consumers more choices in services. With this heightened competition, business owners must accelerate their marketing efforts by deeply understanding consumer needs to deliver precisely targeted products and services. One way to achieve this understanding is by analyzing customer feedback on products and services to refine and improve offerings according to consumer demands.

This data analysis is a key discipline in Data Science, which involves performing Sentiment Analysis on textual feedback to gauge customer attitudes—measuring how satisfied they are with a product—and Classification to categorize the main topics of the feedback.

Given this shift in marketing, the researcher recognizes the importance of analyzing consumer feedback from those who have used the product or service. This enables faster product and service development compared to competitors, allowing the business to grow exponentially. Therefore, feedback from users of the Sriphat Medical Center, Faculty of Medicine, Chiang Mai University, was collected, analyzed, summarized, and presented to determine current consumer satisfaction levels.

The analysis categorizes satisfaction into two levels (Satisfied/Dissatisfied) and identifies the most and least appreciated aspects across eight categories:

1. Service Process
2. Service Behavior
3. Treatment Expertise
4. Food Services
5. Public Relations Information
6. Service Fees
7. Medical Equipment
8. Facility & Environment

This real-time assessment helps the healthcare facility understand how the business is perceived by customers—identifying strengths to promote and weaknesses to urgently address—ensuring continuous growth and advancement.

2 Literature Review

2.1 A Comparison of Similarity Measures for Online Social Media Thai Text Classification

Supatta Viriyavisuthisakul, Parinya Sanguansat, Pisit Charnkeitkong, and Choochart Haruechaiyasak conducted a study to develop a product recommendation system using data sourced from the Pantip website. The data was categorized into four groups. The workflow involved word segmentation using Stop word removal and KUCUT, followed by feature extraction using Term Frequency (TF) and Term Frequency-Inverse

Document Frequency (TF-IDF). Clustering was performed using the K-Nearest Neighbors (K-NN) method. Ten types of distance measures were compared: Bray-Curtis Distance, Euclidean Distance, Minkowski Distance, Cosine Distance, Correlation Distance, Chebyshev Distance, Cityblock Distance, Canberra Distance, Jaccard Distance, and Roger-Tanimoto Distance. The best result was achieved using TF-IDF with K-NN and Bray-Curtis Distance, yielding the highest accuracy at 58.62% [2].

2.2 Foreign trade influence factors research Apply Latent Semantic Analysis to Classify Emotion in Thai Text

Piyatida Inrak and Sukree Sinthupinyo, conducted a study that categorized knowledge from Thai text into six groups. Bi-word analysis was included in the classification. Data collected from the internet was segmented using the SWATH program, focusing only on nouns and verbs. Latent Semantic Analysis (LSA) was used with both single word and bi-word approaches. Words were mapped to emotional categories. Three models were used for emotion classification: Decision Tree, SVM, and Naïve Bayes. Two input variations were tested: single words only, and single words combined with bi-words. The Naïve Bayes model using the combination of single and bi-word inputs achieved the highest accuracy at 90% [3].

2.3 The Comparison of Algorithms for Thai-Sentence Classification by Thanyarat Nomponkrang and Charun Sanrach

Conducted by Thanyarat Nomponkrang and Charun Sanrach, this study aimed to classify Thai sentences into four types: declarative, negative, interrogative, and imperative. Words were segmented and tagged for part of speech using the SWATH program, and stop words were removed. Feature selection methods included term Binary, term Frequency, and TF-IDF. Four models were used for classification: Decision Tree, Naïve Bayes, K-NN, and SVM. The dataset was divided into four groups, including those using key phrases combined with TF-IDF and those using only TF-IDF. The best result was achieved using the SVM model with key phrase extraction and TF-IDF weighted by part of speech [4].

2.4 Combining Lexicon-based and Learning-base Methods for Twitter Sentiment Analysis by Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu and Bing Liu

Conducted by Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu, this study aimed to improve sentiment classification on Twitter. Traditional methods had high precision but low recall, and Twitter text often includes symbols that skew classification. The study addressed these issues by focusing on individual words and incorporating dictionary-based meanings. Twitter data was divided into five datasets. Non-essential elements like retweets were removed, and abbreviations were expanded (e.g., "wknd" to "weekend") using dictionary definitions. Word segmentation and part-of-speech tagging followed. An Augmented Lexicon-based method was used, including sentence classification, sentence linking, and word frequency analysis. Features were selected using Pearson's chi-square test, and classification used Support Vector Machines (SVM). Compared to other methods-ME (state-of-the-art learning-based), FBS (lexicon-based), AFBS (augmented lexicon-based), LLS (AFBS without SVM), and LMS (full process with SVM)—LMS achieved the highest accuracy, recall, and F-score. The LMS method outperformed others due to its broader semantic coverage [5].

3 Data and Methodology

3.1 Data

The feedback data was collected from Sriphat Medical Center, Faculty of Medicine, Chiang Mai University, through various channels including the website, Facebook, Line OA application, and direct phone calls. A total of 6,925 records were gathered. These were divided into:

1. Model development set: 85% (5,886 records)
   1. Training data: 70% (4,120 records)
   2. Validation data: 30% (1,766 records)
2. Blind test set: 15% (1,039 records)

3.2 Methodology

1. Define Sentiment and Topic Classification

The study identified two types of data for analysis:

1. Sentiment classification: Determining if a comment expresses satisfaction or dissatisfaction.
2. Aspect classification: Identifying the main topic of the feedback, classified into 8 categories:
   1. Service behavior
   2. Service process
   3. Food service
   4. Medical expertise
   5. Public relations
   6. Environment & facilities
   7. Service fees

8. Medical equipment
9. Others

2. Data Collection

Permission was obtained to access feedback from multiple platforms as listed above.

3. Data Preparation (Cleaning)

1. Review all collected data by examining each column, prioritizing important information, and remove any unnecessary data.
2. Identify missing data; if essential information is missing, remove those entries.

4. Machine Learning Development

1. Data was loaded into Jupyter Notebook using Python.
2. Sentiment text was processed using the following steps:
    1. Select the "Detail" column, which contains the feedback sentences.
    2. Text Cleaning: Remove symbols, hashtags, and punctuation.
    3. Word Segmentation: Using PyThaiNLP with the 'NEWMM' method, which segments by the longest matching words from a dictionary.
    4. Stopword and Number Removal
3. Feature Extraction: Two methods were used:
    1. Bag-of-Words: Count words that appear more than 25 times (514 words in total).
    2. TF-IDF (Term Frequency-Inverse Document Frequency)
       Use the TF-IDF function, with the parameter *word*, which calculates the weighted word frequency based on document frequency for each token (word or character).
4. Split the data into two sets: a training dataset (Train) and a testing dataset (Test Data).
5. Model Development: Five machine learning models were tested using scikit-learn:
    1. Linear Support Vector Machine (SVM)
    2. K-Nearest Neighbors (KNN)
    3. Stochastic Gradient Descent (SGD)
    4. Naïve Bayes
    5. Decision Tree

Each model's parameters were tuned, and the models were evaluated using the blind test set to ensure the algorithms had not previously encountered the data.

5. Evaluation

Evaluate the performance of the machine learning models by measuring accuracy and analyzing the confusion matrix.

6. System Implementation

1. A web application was developed using Django and Jupyter Notebook for running ML models.

2. Power BI was used for data visualization and integrated into the website for real-time display analysis results.

4. Results

In summary, this research involves the analysis and visualization of sentiment data from consumer feedback in healthcare facilities. The study aims to assess consumer satisfaction with the services provided and identify the aspects that customers are most satisfied or dissatisfied with. The results show that the model using TF-IDF for feature extraction and Linear SVM for classification outperforms other tested models. It achieves 95.8% accuracy in sentiment classification (satisfaction/dissatisfaction) and 77.4% accuracy in categorizing the aspects mentioned in the feedback. The comparison of the model performance is shown in table1 Results of Sentiment Analysis on Evaluation Dataset and Table2 Results of Aspect Classification in Sentences (Evaluation Dataset) below.

However, the model still has some limitations. Errors may occur when processing sentences containing both positive and negative feedback simultaneously. For example:

"การบริการดี พยาบาลน่ารักพูดจาดี แต่ที่จอดรถไม่มีเลย ทำให้ต้องจอดไกล เดินไกลมาก"

Such cases can lead to misclassification. Additionally, the categorization has higher error rates due to insufficient data in certain categories, such as public relations information and medical equipment. Collecting more data in these areas in the future could improve the model's learning and classification accuracy.

The result of Program Evaluation for Feedback Collection and Classification Results

The evaluation was divided into two parts: data visualization and data accuracy. Five experts directly involved in managing and utilizing the feedback data participated in the assessment. They tested the system over two weeks, evaluating both the feedback input module and the results interpretation interface. The findings are as follows:

1. Data Classification Visualization

Users expressed a high level of satisfaction (rating □ 4 out of 5), indicating that the data visualization was user-friendly and met their needs. However, the Word Cloud feature received a moderate rating because it only displayed keywords without detailed context, requiring users to explore further for complete information.

2. Prediction Accuracy

Users rated the prediction accuracy at level 4, indicating that the model's performance was sufficiently reliable for practical use.

Table1 Results of Sentiment Analysis on Evaluation Dataset

Model	Accuracy (Bag of Words Feature Extraction)	Accuracy (TF-IDF Feature Extraction)
Linear, SVM	95.8	95.9
KNN	89.7	73.8
SGD	95.2	95.2
Naïve bayes	87.1	80.4
Decision Tree	92.2	89.2

Table2 Results of Aspect Classification in Sentences (Evaluation Dataset)

Model	Accuracy (Aspect Classification using Bag of Words)	Accuracy (Aspect Classification using TF-IDF
Linear SVM	55.3	77.7
KNN	68	63.5
Linear SVM with Enhanced Efficiency using SGD	70.3	76.5
Naïve bayes	23.8	33.5
Decision Tree	67.4	71.2

5. Comments and Suggestions

1. This research can be further developed to classify satisfaction levels and the aspects mentioned in various healthcare facilities.

2. For system developers, this program can be applied to classify sentiment data, satisfaction levels, and discussed aspects from other sources, such as online reviews on various websites or feedback from other resources like Facebook Messenger and Line OA. This would help cover all feedback channels and reduce the workload of personnel.

3. The model training results revealed that sentiment analysis using the SGD model sometimes misclassified dissatisfaction as satisfaction. This was partly due to frequently occurring words in positive reviews being weighted toward satisfaction, leading to incorrect predictions. The model needs improvement by removing ambiguous words that do not clearly contribute to prediction, such as "เรื่อง" (issue), "จุด" (point), or "คน" (person). Additionally, more training sentences containing positive words but expressing dissatisfaction should be included. Another issue was words with dual meanings depending on context, such as "เร็ว" (fast/too soon), requiring the use of other models that better capture contextual meaning.

4. The aspect classification model showed significant confusion between "service delivery" and "service behavior," as well as between "medical equipment" and "environment." Further analysis identified overlapping high-impact words between these categories, such as "มีข้อบกพร่อง" (flawed), "เสมอ" (always), "ให้ความรู้" (informative), "คำสุภาพ" (polite), "พยาบาล" (nurse), and "โทรศัพท์" (phone) for service-related aspects, and "ปรับปรุง" (improve), "ซ่อมแซม" (repair), "อุปกรณ์" (equipment), and "ความสะดวก" (convenience) for equipment/environment. To improve accuracy, ambiguous words should be removed, and more training data should be added, particularly for underrepresented categories like "environment." Special prediction rules (e.g., classifying "ซ่อมแซม + เครื่องมือ" (repair + tools) as "medical equipment") could also enhance performance.

5. The data visualization component was designed to align closely with user needs. However, it lacked a feature comparing sentiment and discussed aspects over time, making it impossible to immediately identify trends—such as whether negative feedback gradually improved to satisfaction. This aspect should be enhanced to provide comprehensive insights.

6. Users evaluated the sentiment analysis and visualization program's prediction accuracy at an average of 3.8 out of 5 (76%). Follow-up feedback revealed that the system often misclassified sentences based on word frequency alone. For example:

1."พยาบาลที่เคาท์เตอร์พูดไม่ดี พยาบาลคนอื่นก็ยืนนิ่งเห็นดีเห็นงามไปด้วยกันหมด" ("The counter nurse spoke rudely, and others just stood by") was incorrectly labeled as *satisfactory* due to the word "ดี" (good).

2. Conversely, "หมอรักษาละเอียดไม่มีที่ติ รักษาหายจนไม่จำเป็นต้องไปรักษาโรงพยาบาลอื่นต่อ" ("The doctor's thorough, flawless treatment cured me, so I didn't need another hospital") was misclassified as *dissatisfactory* because of "ไม่" (not).

3. Aspect prediction errors also occurred, with key terms like "ขั้นตอน" (procedure) for *service delivery* or "ยิ้มแย้ม" (smiling) for *service behavior* not always triggering correct classification. Refining keyword relevance is needed.

References

1. Cube Analytics Consulting Co., Ltd. "CRISP-DM." [Online]. Available: https://datacubeth.ai/crisp-dm/. [Accessed: March 5, 2021].
2. Kaeroidata. "Web Scraping with Python Using Selenium <Beginner>." *Medium*, 2019. [Online]. Available: https://kaeroidata.medium.com/. [Accessed: May 4, 2021].
3. Pakhapoom Sarapat and Nontawit Cheewaruangroj. "How to Choose the Right Chart for Data Visualization." *bigdata.go.th*, 2021. [Online]. Available: https://bigdata.go.th/big-data-101/picking-chart-for-data-visualization/. [Accessed: July 4, 2021].
4. Julie Zuckerman. "Data Visualization and Visual Analytics: Seeing the World of Data." *Sisense*. [Online]. Available: https://www.sisense.com/blog/data-visualization-and-visual-analytics-seeing-the-world-of-data/. [Accessed: July 5, 2021].
5. Saksit Srimarong. "Visual Analytics: Data Analysis Through Visualization." *bigdata.go.th*, 2021. [Online]. Available: https://bigdata.go.th/big-data-101/visual-analytics/. [Accessed: July 5, 2021].