SENTIMENT ATTRIBUTION ANALYSIS WITH HIERARCHICAL CLASSIFICATION AND AUTOMATIC ASPECT CATEGORIZATION ON ONLINE USER REVIEWS

Myat Noe Win[1], Sri Devi Ravana[2*] and Liyana Shuib[3]

[1,2,3]Department of Information Systems, Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

Email: myatnoewin786@gmail.com[1], sdevi@um.edu.my[2*](corresponding author), liyanashuib@um.edu.my[3]

DOI: https://doi.org/10.22452/mjcs.vol35no2.1

ABSTRACT

Due to COVID-19 pandemic, most physical business transactions were pushed online. Online reviews became an excellent source for sentiment analysis to determine a customer's sentiment about a business. This insight is valuable asset for businesses, especially for tourism sector, to be harnessed for business intelligence and craft new marketing strategies. However, traditional *sentiment analysis with flat classification and manual aspect categorization technique imposes challenges with non-opinionated reviews and outdated pre-defined aspect categories which limits businesses to filter relevant opinionated reviews and learn new aspects from reviews itself for aspect-based sentiment analysis. Therefore, this paper proposes sentiment attribution analysis with hierarchical classification and automatic aspect categorization to improve the social listening for diligent marketing and recommend potential business optimization to revive the business from surviving to thriving after this pandemic. Hierarchical classification is proposed using hybrid approach. While automatic aspect categorization is constructed with semantic similarity clustering and applied enhanced topic modelling on opinionated reviews. Experimental results on two real-world datasets from two different industries, Airline and Hotel, shows that the sentiment analysis with hierarchical classification outperforms the classification accuracy with a good F1-score compared to baseline papers. Automatic aspect categorization was found to be able to unhide the sentiment of the aspects which was not recognized in manual aspect categorization. Although it is accepted that the effectiveness of aspect-based sentiment analysis on flat classification and manual aspect categorization, none have assessed the effectiveness while using hierarchical classification with a hybrid approach and automatic aspect categorization.*

Keywords: Sentiment Analysis, Online User Reviews, Hierarchical Classification, Automatic Aspect Categorization, Topic Modelling, Topic Cluster Labelling, Business Optimization

1.0 INTRODUCTION

When a service is delivered by the business, the customer is either satisfied or dissatisfied. This sentiment of the customer plays an important role in his or her future purchasing decision. Customers' reviews on the purchase's items, as a means of word-of-mouth communication, are shoppers' voluntary input on their buying experience, and they can provide a lot of valuable insights to both customers and businesses [1]. Nowadays, with the popularity of social networking, customers also take their experience straight to the internet. According to the Customer Review Trends 2020 by BrightLocal, 93% of consumers browsed online for a local business, and 31% of consumers claimed that they read online reviews more in 2020 due to COVID-19 [2]. When a favorable experience on the service is shared by the user through a positive sentiment review, that will inspire others to purchase from that business. In 2020, 79% of consumers claimed that they trust online reviews similar to the personal recommendations from the close circle. Besides, 94% of online shoppers incorporate reviews into their purchase decision, and 92% of people will hesitate to purchase from a business that has negative online reviews [2]. As a result, consumers' evaluations of brands, purchasing choices, and purchase habits will all be influenced by online feedback of bought items [1].

The above statistic from BrightLocal indicated the impact of sentiment on the purchase decision and the pivotal of sentiment analysis on the user reviews, i.e., a process to determine the positive and negative sentiment of the user on a particular service or product thru various data analysis approaches. In response to the growing and available online user review data, academic research has also shifted its emphasis to online reviews and prioritized in the area of emotional tendency, emotional polarity, and emotional categorization on the sentiment in online feedback [3][4]. Therefore, online user reviews are heavily used in the sentiment analysis to understand the sentiment towards the provided services to align strategic direction to achieve maximum customer satisfaction and improve business service quality.

Out of all the existing mature markets such as e-commerce and retails, this paper focuses on online user reviews from the airline and hotel domain as the domain of interest, because these two domains are significantly impacted by the COVID-19 pandemic and required implementable business optimization opportunities to recover the experienced loss.

1.1 Importance of Sentiment analysis in Airline and Hotel Domain

In this context, airline as a service-oriented business sector is on-demand for sentiment analysis to understand their customers' sentiment and the correlation to purchase decision. In fact, airline business performance is steered by Airline Customer Satisfaction on Airline Service Quality [5]. It is crucial for airlines to understand the strength and weakness of their service quality, as well, their customers' sentiment on it. After the COVID-19 pandemic and tourism crisis, aviation sectors may become more customer-centric and competitive for market leadership [6]. Therefore, airlines need to constantly analyze online customer reviews to provide a better quality of service and achieve customer satisfaction, which will assist the airlines in leading a competitive advantage [7]. Hence, with the rise of open data initiatives and the increasing amount of online user reviews data available, there is a greater need for an optimized approach for sentiment analysis to provide more reliable results with greater efficiency and effectiveness as per business needs of airlines.

Similarly, the Hotel industry is one of the dominant industries in which consumers are reading reviews before making any purchasing decision. 87% of the consumers believe that reviews play an important role in this industry [2]. Henceforth, there is a demand to understand the customers' sentiments and purchasing decisions in these industries, which will determine the facets that fuel the business optimization before resuming business-as-usual after the COVID-19 pandemic.

1.2 Sentiment analysis and Aspect-based Sentiment Analysis

In a nutshell, the science behind the sentiment analysis is the process of categorizing customers ‘sentiments into positive, neutral, or negative, through the different data analysis approaches. Sentiment analysis can be achieved by various types of algorithms under three main approaches: Statistical approach, knowledge-based approach, and hybrid approach [8]. In a standard sentiment analysis process, after collecting online user reviews, the dataset is cleaned, labelled, and directly fed into the chosen approach or algorithm to perform sentiment classification, and this process is widely known as flat classification. As of the current state-of-arts, sentiment analysis is performed mostly on flat classification, by directly feeding the cleaned dataset into sentiment classification without subjectivity classification i.e., without filtering the non-opinionated or objective or factual texts. As a result, the dataset will unnecessarily grow with non-opinionated texts, which is ineffectual for the sentiment analysis process, and adding challenge to modelling or classification. As a result, the high number of the neutral polarity may bias the dataset or overfeed the model. With this practice, authors highlighted that one of the major issues in the current sentiment analysis approach is incorrectly classification of the major portion of the opinionated data into a neutral class or vice versa, due to lack of subjectivity classification prior to sentiment classification. Authors also claimed that detection of opinionated texts (subjectivity classification) prior to sentiment classification is a simple process relevant to all business domains but lacking in the most sentiment classification process [8]. Likewise, authors of [9] concluded that flat classification might feed unnecessarily big data to the model and might bias the accuracy of the model in the classification process.

On the other hand, aspect-based sentiment analysis (ABSA) is a sub-technique under sentiment analysis that will highlight the customer sentiment on a particular aspect of the service from the user reviews and assists in exploring a new market, anticipating future trends, or having the edge over the competition [10]. For ABSA, various aspects of the service are needed to extract from the user reviews to map the aspect and the sentiment of the user on that particular aspect. Generally, in ABSA, aspect extraction is performed by capturing and sorting of high-frequency words (HFW) from the user reviews and mapping of extracted aspects or HFW to the predefined aspect category list obtained from the domain-related portals or previous work, and this process is known as manual aspect categorization in ABSA. As of the current state of arts, manual aspect grouping is widely used in aspect categorization, and taxonomizing is limited to predefined aspect list and prunes to human errors [11], also manual verification is required to unify similar words [12]. Consequently, manual aspect categorization costs expensive effort and produces reliability issues due to manual synonym unification in aspect list, challenges to capture the new aspects from new review comments, aspect-mapping limited to pre-defined list, inability to discover the meaningful aspects from the review itself, and high tendency to rely on outdated pre-defined list. However, new features are introduced in business, reviews are generated every day and our sentiment changes over time on different dimensions of the products. Therefore, the aspect categories validated and pre-defined previously may not be sufficient.

In brief, regardless of the chosen approach, the generic theoretical concept of sentiment analysis solely relies on the

opinionated texts in the reviews because every user review contains two types of texts: 1) opinionated or subjective texts and 2) non-opinionated or objective texts. Therefore, the quality of filtering opinionated texts from the non-opinionated texts is an important process in sentiment analysis to perform analysis on the most relevant data. Similarly, the result of ABSA solely relies on the aspects being extracted and categorized from the user reviews. The elimination of the aspects or overlooking of the aspect category might produce an incorrect mapping of aspect to sentiment and an unreliable end result for the ABSA. Henceforth, as far as the previous works are aware of the importance of classification and categorization in sentiment attribution analysis, there is no work carried out on hierarchical classification to overcome the limitation of flat classification where current state-of-art have not comprehensively considered the optimum technique to eliminate the non-opinionated texts from the user reviews prior to sentiment classification. Likewise, according to the author's knowledge, no comprehensive work was dedicated to automatically build aspect categories from the review itself to overcome the limitation with outdated aspect categories contributing to inaccurate aspect-based sentiment analysis.

Henceforth, the objective of this paper is to perform sentiment analysis on online user reviews with hierarchical classification and automatic aspect categorization to unhidden the business insights from the review data itself and recommend the aspects that drive the business optimization. The motivation of this paper is to introduce the practical implication from the research, which can contribute to societal benefit. Therefore, it presents a solution on the applicability of technology to understand the psychological perspective of individuals and assist in business excellence, i.e., using information retrieval (technology) to understand the sentiment (psychological perspective) from online user reviews and advance the business operations.

Therefore, this paper aims to propose a new approach in classification and categorization for sentiment attribution analysis by introducing; 1) hierarchical classification technique using a hybrid approach (rule-based and machine learning) to build an opinionated review dataset for sentiment classification and 2) automatic aspect categorization technique to unhidden the aspects and build aspect categories from the online user reviews itself. Moreover, the paper brings a novelty on the incorporation of hierarchical classification (using hybrid approach) in Sentiment Analysis instead of Flat Classification, incorporation of Automatic aspect categorization in Aspect Extraction instead of Manual Aspect Categorization and incorporation of Topic Modelling on opinionated review comments for ABSA. Additionally, from the practical implication, the outcome of this research will add an advantage to business service owners to understand the customer's sentiment on their as-is services to optimize their business operations and to recover their loss during this pandemic. From the perspective of customers, the outcome of this research will prove that it can filter the relevant opinionated reviews of users from the massive users' reviews and capture the actual sentiments to better understand the customer's point of view in the business services. This will ensure that the reviews provided by the users are not wasted or overlooked. All in all, this research will lead to new research avenues in the area of hierarchical sentiment classification and automatic aspect categorization in ABSA.

This paper is divided into six sections. The first section is the introduction of the research. Second section is the summary of current state-of-art approaches and the motivation of choosing the proposed approach. Section 3 will discuss the research methodologies employed during this research. The data collection and the origination of the data sets will be explained. Section 4 will explain the experiment setup and the results. The findings from the experiment will be presented in section 5 followed by the conclusions reached by the research study in section 6.

2.0 BACKGROUND

There are state-of-art studies focusing on the latest approaches and performance measuring methods and limitations in flat classification and aspect categorization. In Hotel and Airline domains, the previous research focuses primarily on the flat classification and manual aspect categorization using supervised machine learning models, dictionary-based approaches, matrix decomposition approaches and commercial tools available in markets.

2.1 Sentiment analysis on flat classification

Authors of [9] performed sentiment analysis on Twitter US airline review dataset using different classification strategies such as Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors, Logistic Regression, Gaussian Naïve Bayes and AdaBoost. The result concluded that the classification with AdaBoost gives an accuracy of 84.5%. However, the classification is directly performed on the cleansed dataset as a flat classification on the limited number of tweets. Authors claimed that a dataset with neutral polarities in the training set might feed unnecessarily big data to the model and might bias the accuracy of the model without proper up sampling.

Similarly, authors of [13] used Naïve Bayes (NBC), Support Vector Machine (SVM) and Radom Forest to conduct

the sentiment analysis on Philippine's airlines reviews from Twitter. The authors claimed that NBC is the closest in recognizing the true polarity of the tweet based on the average of its accuracy and kappa. In this paper, the authors acknowledged the importance of the correct labelling of the reviews prior to the sentiment classification and highlighted that dataset labelling is fully done manually by three annotators to improve the correctness of the labelling. However, it is observed that neutral polarity reviews in the dataset contributed to decrease of both the accuracy and kappa, while it is directly fed into the classification algorithm.

Likewise, authors of [14] performed human labelling on ambiguous tweets and conducted sentiment classification using Naïve Bayes (NBC) algorithm. The result of the experiment was able to classify ambiguous tweets and neutralize them according to their weight, with 0.79 as worst and 0.90 as best f-score. However, it is observed that classification without subjectivity detection might increase the noise of irrelevant tweets.

2.2 Sentiment analysis with subjectivity classification

On the other hand, observing the importance of filtering the opinionated texts from the review dataset, few authors have introduced a rule-based approach for subjectivity classification prior to topic modelling in sentiment analysis.

Authors of [15] performed sentiment classification using NBC with Laplace estimator on Massive Open Online Course (MOOC) course review data from the class-central.com. The paper stated that the limitation in this paper is NBC does not get the necessary information about the minority class to build an accurate prediction. Therefore, there is an imbalance distribution in precision and recall score, also the low weighted value of F-measure. Thus, the authors suggested using some mechanism for sampling method to modify imbalanced data into balanced distribution [15]. In 2019, authors of [16] performed sentiment classification with Lexicon Bing for subjectivity classification and LDA for topic modelling. However, the approach only used a lexicon-based approach for subjectivity classification, which might not be reliable for subjectivity detection as the biggest downside of the rule-based method is that the procedure only cares about individual words and lacks the context in which it is used [16]. With noisy texts such as online user feedback, such an approach can easily be misunderstood [17]. Additionally, the paper stated future work to evaluate the performance with coherence score, instead of perplexity score alone.

Likewise, another research on subjectivity classification using a rule-based approach is conducted on the Twitter dataset. In this research, a lexicon-based approach for subjectivity detection with a dictionary and corpus-based methods (SentiWordNet, Word-based corpus, Context-based corpus) is used. The result is evaluated using three different datasets. The experimental results have also been compared with expert opinions to evaluate the significance of the proposed lexicon-based subjectivity detection sentiment analyzer [18]. However, this approach is solely a lexicon-based approach for classification. It might not be reliable for subjectivity detection because the rule-based approach cannot handle the context and only recognize the keyword and ignore the rest of the sentence. Therefore, the result of this subjectivity classification will not guarantee the reliable opinionated reviews, and further screening will be needed to filter the more accurate opinionated reviews from the dataset.

2.3 Aspect Categorization in Aspect Based Sentiment Analysis

On the other hand, for ABSA and aspect categorization, [28] performed ABSA on customer reviews data of Soekarno-Hatta Airport using NBC and SVM for flat classification and manual aspect taxonomizing by Skytrax Website for aspect categorization, on the aspects extracted by Matrix decomposition (Term frequency-inverse document frequency (TF-IDF), Singular Value Decomposition (SVD)). The result discussed each aspect of the airport as found in the aspect category (whereas the aspect list is limited by the aspect categories on the Skytrax website).

Moreover, authors of [19] conducted ABSA using Pearson correlation of airline features with an overall rating from Skytrax portal and manually adding sentiment using AlchemyAPI for aspect categorization. Sentiment classification was performed as flat classification with NBC, C4.5, Random Forest, CART, Hoeffding Tree algorithm. The authors stated that Hoeffding Tree provides strong accuracy and a competitive runtime with F1 = 0.839 for review text sentiment [19]. However, the aspect for each service is determined in the scope of the Skytrax portal as aspect categorization is performed manual mapping with a predefined list.

Also, sentiment analysis on Meituan.com online hotel reviews [12] used TF-IDF to extract the top 20 features of all comments and manually eliminating and unified high-frequency words from the top 20 to form the hotel feature list. The proposed model can only extract features and viewpoints in explicit sentences. The author stated that manual effort is needed for synonym clustering using manual verification of similar meaning in the aspect list.

Not restricted to only airline domain, this limitation is found in other domain areas as well, for example, sentiment analysis on online Zomato customer reviews on the Indonesian restaurants is performed with flat classification on unsupervised learning that performs aspect extraction, aspect sentiment orientation classification and Manual aspect labelling reference to [20] and [21] to create aspect list with predefined four aspect labels [22].

Authors of [23] used AYLIEN Text API built on hierarchical Bidirectional Long- and Short-Term Memory (HLSTM) for the Yelp restaurant review dataset. The author stated that the result returned a 39.90% accuracy rate, 60.7% error rate, but AYLIEN Text Analysis API is paid version, and the accuracy rate is not reliable enough for aspect categorization. Additionally, similar to AYLIEN Text API, there are few other well-known commercial sentiment analysis products available in the market such as IBM Watson's Natural Language Understanding APIs and Tone Analyzer. However, the same limitation imposed as paid features are limited for testing and comparison.

Likewise, sentiment analysis on Amazon product review was performed through flat classification with NBC and measure the result with precision, recall and f-measure. Also, LDA topic modelling is used to extract topics and topics are mapped to aspects identified from the product descriptions such as Network, Battery, Price, etc. The authors mentioned developing a topic-aspect mapping algorithm to have aspect extraction from datasets instead of predefined categories as a part of future work [11].

2.4 Research Gap and Findings

In brief, little work has been carried out on subjectivity classification, and previous works have not comprehensively considered hybrid approaches in hierarchical classification. The latest studies concluded a knowledge gap for an improved hierarchical classification approach as the current classification is either flat classification or lexicon-based subjectivity classification. Thus, the limitations of each of these approaches will hinder the effectiveness of sentiment classification. Indeed, at present, general research in hierarchical classification is still in its infancy. Recently, [37] conducted a research on hierarchical classification using a rule-based approach on Twitter dataset. In this research, lexicon-based approach for subjectivity detection is used with a dictionary and corpus-based methods (SentiWordNet, Word-based corpus, Context-based corpus). The result is evaluated using three different datasets, and the experimental results have also been compared with expert opinions to evaluate the significance of the proposed lexicon-based subjectivity detection sentiment analyzer. However, this approach is solely a lexicon-based approach for classification and might not be reliable for subjectivity detection because the rule-based approach cannot handle the context and only recognize the keyword and ignore the rest of the sentence.

Therefore, this paper will introduce hierarchical classification using a hybrid approach (rule-based and machine learning) to build an opinionated review dataset without objective texts or neutral polarity in order to perform more reliable sentiment classification, especially with the noisy texts from review comments. The hybrid approach will markup the deficiency in each approach and produce a better result as rule-based will decrease the impact of human labelling mistakes in classification input, and the machine learning approach will improve the context-based classification in subjectivity detection.

Similarly, although aspect extraction is well established for aspect-based sentiment analysis, its effect on automatic aspect categorization from the review itself has not been extensively studied. The latest studies summarized that current aspect categorization is solely depended on the manual aspect list or pre-defined aspect list. Therefore, this paper will introduce the automatic aspect categorization to automatically identify the aspects from the user review and build aspect categories to perform topic-aspect mapping in ABSA. The result will unhide the valuable aspects from the reviews which were not listed in the predefined aspect category list in available review platforms.

3.0 DATA AND METHODS

This study adopts a quantitative research design with experimental analysis, and the overall workflow for the proposed approach is demonstrated as below in Figure 1 and explained further in section 3.1. to 3.7.

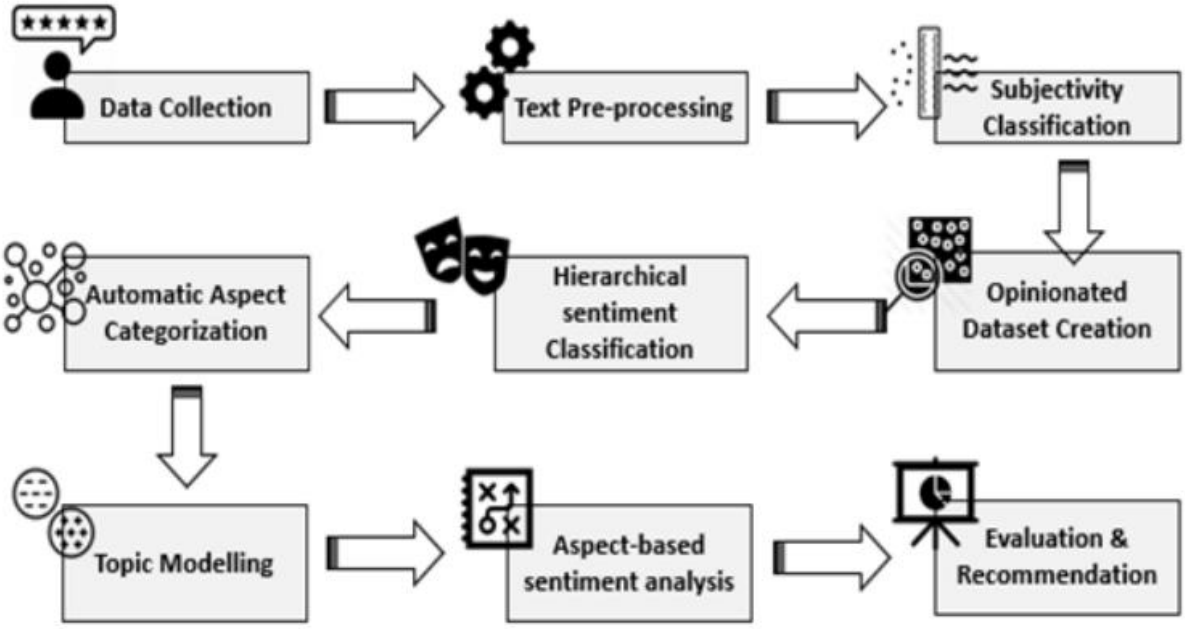

Fig 1: Proposed framework for Sentiment Analysis for Business Optimization

3.1 Data Collection

There are two datasets used in this paper, i.e., the Airline user review dataset and the Hotel user review dataset. The dataset collected is primary data with random sampling. For the airline dataset, 65947 online user reviews from the Skytrax airline review portal for 81 airlines around the world in English corpus for the period of 2002 to 2019 [24]. For the hotel dataset, 10000 reviews about 1,400 hotels from Dec 2018 and May 2019 in English corpus is collected from Kaggle [25]. The user review datasets from these two domains are chosen because their business operation is highly impacted by the COVID-19 pandemic and the purpose of this research is to provide practical implication on loss recovery for the impacted sector after this pandemic. Additionally, purpose of using the existing datasets is to identify whether we can capture new aspect categories from these existing reviews, by applying the proposed technique, which was not able to capture by the state-of-art studies using the same datasets.

3.2 Text Pre-processing

Data cleansing is performed using traditional state-of-art methods and customized cleaning functions. The traditional cleaning process consists of removing stop-words, non-letters and non-spaces, numbers, punctuation, whitespaces, double spaces, transforming to lower-cases, lemmatization and stemming, and filtering tokens (by the length of word less than 3). Customized cleaning includes removing abbreviation, flight path abbreviation, unrecognized words and autocorrecting the misspelled words using the Symspell package. After intensive cleaning process with testing of different cleaning methods iteratively to improve topics, sentence splitting is performed on the reviews based on the four "negation words" (i.e., "but", "yet", "however", "instead"). Moreover, the attributes for positive and negative words are classified with part-of-speech tagging where sentiment-rich words are often adjectives and correlated with sentiment polarity (positive or negative).

3.3 Subjectivity Classification and Opinionated Dataset Creation

Subjectivity classification is the process of filtering factual reviews or neutral polarity scored reviews from the dataset and constructing a helpful review dataset on the subjective dataset. Subjectivity classification is performed with three steps with the combination of rule-based and machine learning as hybrid, i.e., first-level subjectivity detection with TextBlob, second-level neutral polarity elimination with Vader and subjective dataset validation with Flair. Firstly, the dataset is cleaned and run TextBlob to capture the subjective comments, i.e., text with subjective polarity > 0.1 and neutral polarity analysis with sentiment polarity between -0.1 and 0.1. For the TextBlob subjectivity detection, the subjective value is calculated between [0.0, 1.0]. The subjectivity score for the given dataset will be as in Equation 1 and Equation 2.

$$R_{S_{(i)}} = T_{s_{(i)}} \geq 0.1 \quad (1)$$

$$R_{o_{(i)}} = T_{s_{(i)}} < 0.1 \quad (2)$$

Where:

- T = TextBlob
- R = User review
- s = Subjective
- = Objective
- i=Number of occurrences

The output dataset is input for Vader neutral polarity detection, i.e., text with sentiment compound score of between -0.05 and 0.05 to cover the social media text analysis. On the other hand, for the Vader, the polarity is measured with the compound score as below in Equation 3 and Equation 4.

$$R_{n_{(i)}} = -0.05 < V_{c_{(i)}} < 0.05 \quad (3)$$

$$R_{nn_{(i)}} = V_{c_{(i)}} \geq 0.05 \;||\; V_{c_{(i)}} \leq -0.05 \quad (4)$$

Where:

- V = Vader
- c = Compound score
- R = User review
- n = Neutral
- nn = Non-neutral

For the rule-based models, Vader (i.e., a parsimonious rule-based model for sentiment analysis of social media text) outperforms TextBlob (i.e., simple rule-based API for sentiment analysis on SentiWordNet) because it was developed with an emphasis on the texts on social media which capture the essence of texts that are characteristic of social media – short phrases with emojis, repetitive vocabulary and copious use of punctuation (such as exclamation marks) [26]. Thus, in this research, TextBlob is chosen for subjectivity score for first level subjectivity screening and Vader for second-level neutral polarity detection.

After that, Flair is used on the Vader output to validate the sentiment polarity with a score calculated for each sentence and the score < 0 is negative while the score > 0 is positive to capture the sarcasm and ambiguous sentences. For example, "the food tastes so delicious as pet food" will be positive when considered by any token-based algorithms. Therefore, in this research, Flair (as a machine learning classification model) is used to perform the polarity validation and user label validation, i.e., to check if there is any inconsistency between the scores of Flair versus Vader versus User label. A few tests have been run to validate the ability of Flair on context-based classification for the reviews outperforms the Vader. Henceforth, three steps are taken with a hybrid approach to markup the deficiency of each approach and bring the nearest accuracy for subjective classification.

On the other hand, opinionated dataset creation is the process of re-validating the subjective dataset with the opinionated lexicon to only collect the opinionated reviews from the dataset. In this step, opinionated review extraction is performed with the designed algorithm to check the appearance of positive and negative opinion words (from opinion-lexicon-English) [27] in each review and extract opinionated reviews based on the match. The outcome of this final process creates the opinionated airline review datasets with only helpful subjective reviews.

3.4 Hierarchical Sentiment Classification

Hierarchical sentiment classification is the process of performing sentiment classification as level 2 on top of subjectivity detection (i.e., level 1), as shown in Figure 2. The opinionated airline review dataset as the result of subjectivity classification (described in 3.3.) is used as an input dataset for level 2.

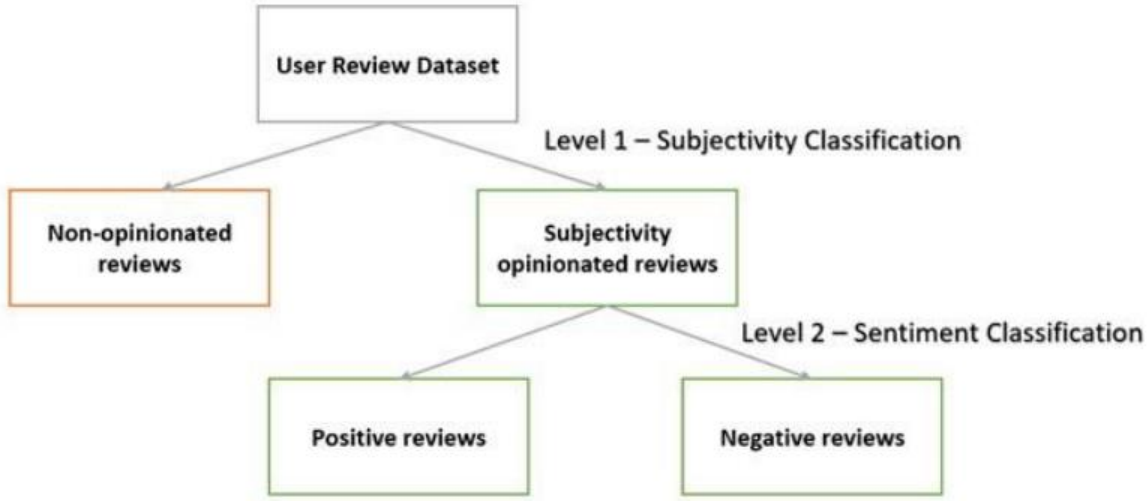

Fig 2: Proposed Hierarchical Classification Approach

After analyzing the dataset, this study proposes an effective labelling approach for the chosen dataset, i.e., initial sentiment score based on the overall rating and recommended rating columns in the dataset, then second-level sentiment score calculation using a combinational approach in probability. The dataset comes with linear 10 scale for overall rating as direct feedback extraction from users. Therefore, for overall rating, the rating and corresponding sentiment assignment is reviewed through multiple literature to make the conclusion that 10 = Perfect/Masterpiece, 9 = Superb/Almost Perfect, 8 = Great, 7 = Good, 6 = Average/Passable, 5 = Mediocre/Unsure, 4 = Poor, 3 = Bad, 2 = Awful and 1 = Worst. Based on these criteria, sentiment polarity is divided accordingly. This labelling is significant as existing research on the same dataset (i.e., Skytrax online user reviews) labelled by using either overall rating column (1-10) or recommended column (Yes/No) to label the dataset [28][19]. The limitation with that is the possibility to skip the values that are not assigned to the overall rating or recommendation, and those reviews might get eliminated. Finally, the labelled dataset is validated with the eight eyes principle where four dedicated annotators (from Psychological, educational background) reviewed the content of the correctness and signed off the subjectivity dataset for further experiment in this study.

After pre-processing, the data is split into 80% (Training) and 20% (Testing) using the Stratified sampling and Synthetic Minority Oversampling Technique (SMOTE) to handle the selection bias or imbalance data. Stratified sampling builds random subsets where each subset contains approximately the same proportions of the class labels. On the other hand, SMOTE will balance class distribution by randomly increasing minority classes by replicating them. Also, 10-fold Cross-Validation is used to handle the bias and variance, i.e., every data point gets to be in a validation set exactly once and gets to be in a training set k-1 time. On the other hand, for the supervised machine learning models for sentiment classification, the linear classification model (SVM) and probabilistic classification model (NBC) is chosen as state-of-art approaches in sentiment analysis [13]. NBC is a simple classifier that classifies based on probabilities of events based on prior knowledge of conditions that might be related to the event. It is an independent cluster and fast in computation to use even in real-time sentiment analysis. On the other hand, SVM is the linear classifier and can handle large scale data. Most text classification problems are linearly separable. In the text classification, both the numbers of instances (document) and features (words) are large. The linear kernel is good when there is a lot of features (TF-IDF result).

Algorithm for NBC:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

- P(A | B) =Probability of A occurring given evidence B has already occurred
- P(B | A) =Probability of B occurring given evidence A has already occurred
- P(A) =Probability of A occurring
- P(B) =Probability of B
-

Algorithm for linear kernel:

$f(x) = B(0) + sum\ (ai * (x, xi))$

Where:

- f(x) = Prediction for a new input using the dot product
- x,a = Support vector input
- B(0) = coefficients

3.5 Automatic aspect categorization and Topic Modelling

Aspect-based sentiment analysis is more customer-centric than topic-based sentiment analysis when it comes to assessing consumer emotion. ABSA provides a comprehensive perspective of consumer preferences, allowing to focus on the specifics of which elements of an organization can be improved. Topic analysis, on the other hand, provides a broad overview of the customer's overall impressions of the product or service [29]. Therefore, incorporation of the topic modelling approach to aspect categorization produces a better result in terms of competitive advantage exploration in sentiment attribution analysis. Therefore, online user reviews are processed into opinionated reviews and performs sematic similarity clustering and LDA topic modelling to unhidden the meaningful aspects from the opinionated review dataset, as shown in Figure 3.

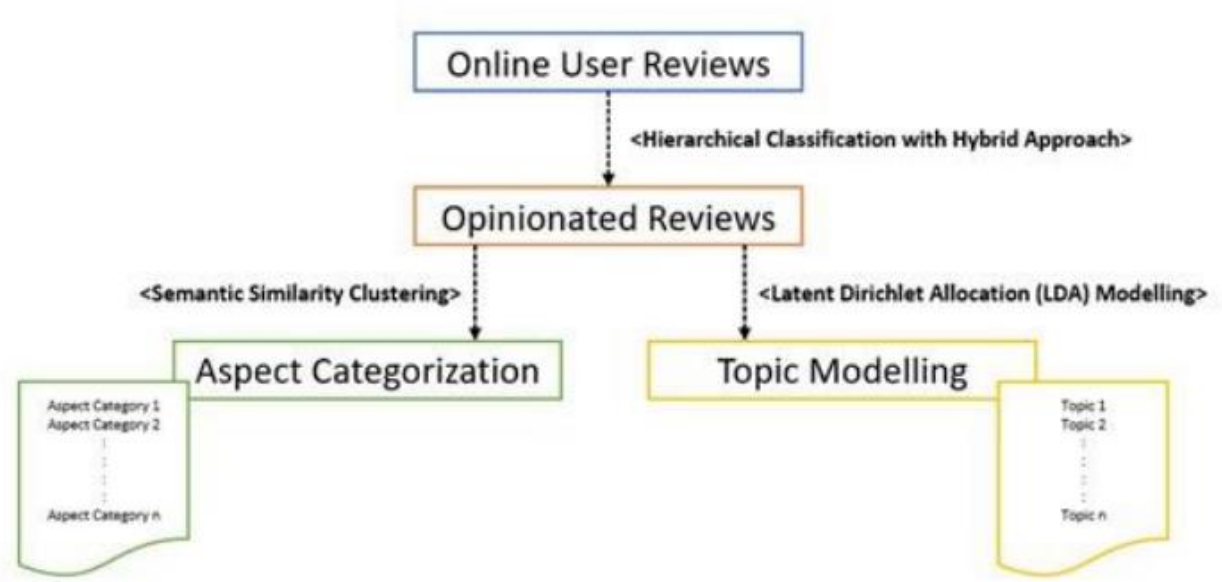

Fig 3: Framework for Aspect Categorization and Topic Modelling

Hence, this paper proposed unsupervised machine learning for semantic similarity clustering using GloVe (Global Vectors for Word Representation), K-means clustering and LDA for topic modelling to understand the correlation between the elements in the review and construct the aspect categorization and topic aspect mapping on it. For aspect categorization, steps consist of extracting the high-frequency nouns using TF-IDF, converting it to word vector representations using GloVe, inputting the vector to k-means for clustering, clustering similarities and building the aspect category list from the review itself. Mapping logic is designed for the aspects (elements of k-means clusters) to aspect categories by identifying the words in a particular cluster, to check the number of words and correlation in one cluster compared to the words in real-life feature dictionary, for example, Aspect Category 1 (aspect 1,.., aspect n), Aspect Category 2 (aspect 1,.., aspect n), etc. Multiple clusters can belong to one aspect category because there are different dimensions under one category, e.g., geographical related cluster, flight arrival/departure destination clusters can belong to flight route.

Algorithm for TF-IDF:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)
IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

On the other hand, LDA topic modelling is performed on the opinionated dataset with an LDA mallet. It builds a topic per document model and words per topic model, modelled as Dirichlet distributions. The different random seed is

used to find the best coherence value for the number of topics and detect the optimal number of topics for the dataset. LDA generates distinct topics and list of topic keywords as Topic 1 (keyword 1,.., keyword n), Topic 2 (keyword 1,..keyword n), etc.

Algorithm of LDA mallet:

$$p(word\ w\ with\ topic\ t) = p(topic\ t \mid document\ d) \times p(word\ w \mid topic\ t)$$

Where:

- p(topic t | document d) = the proportion of words in document d that are assigned to topic t
- p(word w| topic t) = the proportion of assignments to topic t over all documents that come from this word w
- p(word w with topic t) = Update the probability for the word w belonging to topic t

For topic-aspect mapping, each keyword from the topic will be looked up with the aspect category based on the aspect list from k-means clusters to collect aspect-keyword relations. Then, each topic is mapped to a particular aspect category based on the number of words in topic keywords that are related to an aspect category. Reviews are mapped to Topics, and Topics are mapped to Aspect Categories. Topic cluster labelling is achieved by mapping keywords from topics to aspect lists from aspects, as shown in Figure 4.

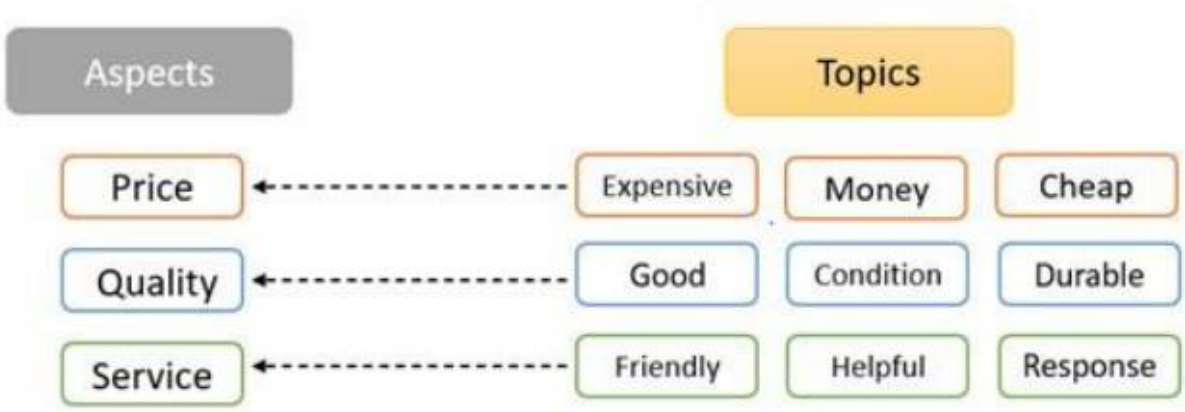

Fig 4: Framework for Aspect-Topic Mapping

For topic-review mapping, designed algorithm lookup opinionated reviews to find topics in each sentence and extract aspect specific reviews. Each of the user reviews is scanned to check if it contained topic keywords. That particular user review is assigned to the corresponding aspect category based on the dominant topic percentage relevance using LDA mallet.

3.6 Aspect-based sentiment analysis and evaluation

After defining the topic-aspect relation, the individual aspect category is mapped to the corresponding predicted sentiment class (positive or negative) based on the number of terms in the topic that is related to an aspect and sentiment. Each of the aspect specific reviews has polarity assigned to it, and that determines the sentiment of that review. Therefore, each aspect or topic has its polarity assigned to it. The number of positive reviews and negative reviews for each topic that is related to the aspect category is analyzed for each user review. If the total positive sentiment polarity is higher than the negative, this particular aspect is reviewed as positive. In contrast, if the total negative sentiment is greater than the positive, then the sentiment of that particular aspect is reviewed as negative. The aforementioned process is illustrated as in Figure 5 to highlight the steps in the framework that drive the summary of aspect-based opinion in proposed sentiment analysis.

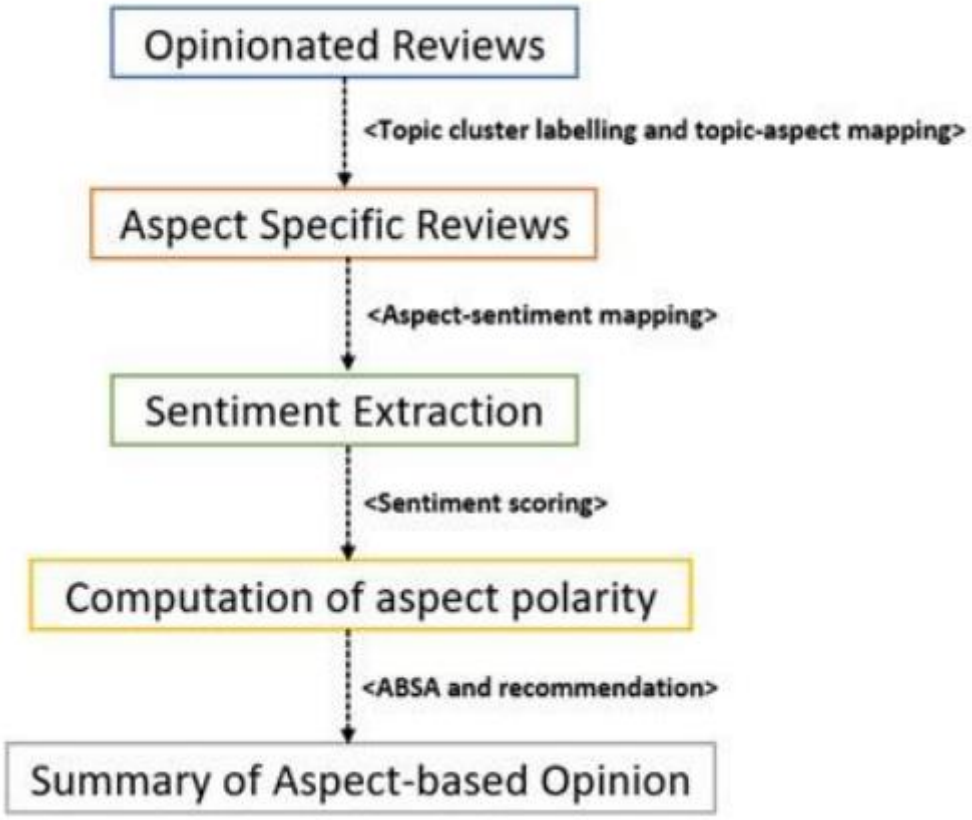

Fig 5: Framework for Aspect Based Sentiment Analysis

All in all, the result of the ABSA will highlight the aspects that steer the business performance of the airlines and customer sentiment influence on each of the aspect categories. Thus, based on the probability, topics are mapped to the corresponding aspects, and then polarity is computed at aspect-level, then classification of sentiments is performed using NBC and SVM.

3.7 Evaluation and recommendation

Different baselines are set for the evaluation of the proposed approach with unit evaluation and integration evaluation. For classification, performance evaluation is measured using Precision, Recall, F-measure and F-measure is used as the main metric to evaluate the model with both good precision and recall and comparison with baseline papers using the same evaluation metric. Precision, Recall and F-measure are three factors in effective sentiment analysis evaluation [30]. Precision is a measure of how often a sentiment rating was correct. Recall is a measure of how many comments with sentiment were rated as sentimental. F-measure is a holistic account of overall performance with a balance between precision and recall for the classifier.
The formula of F-measure [30]:

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

Once the model's performance is confirmed to outperform the existing baseline papers and approaches, another round of evaluation is conducted on the proposed model using Receiver Operation Curve (ROC) to ensure the stability of model performance with different performance evaluation criteria. ROC summarize the trade-off between the true positive rate (tpr) and false positive rate (fpr) for a predictive model using different probability thresholds. It has been recognized as the higher the AUC, the better the performance of the model [31]. When AUC is above 0.5, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values.

For topic modelling, a coherence score is used to evaluate the performance of the topic extraction. coherence c_v is used to evaluate the quality of the learned topics, the closer to 1, the better. c_v is based on a sliding window, a one-

set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosinus similarity [32].

Additionally, the result of the aspect extraction is validated with real-life business review articles. Besides, the models implemented in this research is hyper-tuned to achieve the desire evaluation result before providing the recommendation based on the analysis result. The final visualization and analysis of the ABSA result recommend the action items for the Airline and Hotel domains in order to leverage the business performance.

4.0 RESULTS

The outcome of the hierarchical classification experiment is shown in Table 1.

Table 1: The result of the Hierarchical classification experiment for Airline Domain

Experiments	Precision	Recall	F-measure
Exp1.1: NBC on Flat Classification for Airline domain	45.92%	82.44%	**58.98%**
Exp1.2: SVM on Flat Classification for Airline domain	65.83%	75.44%	**70.31%**
Exp1.3: NBC on Hierarchical Classification (Rule-based) for Airline domain	57.8%	87.2%	**69.52%**
Exp1.4: SVM on Hierarchical Classification (Rule-based) for Airline domain	76.95%	76.59%	**76.77%**
Exp1.5: NBC on Hierarchical Classification (Hybrid) for Airline domain	85.06%	83.72%	**84.39%**
Exp1.6: SVM on Hierarchical Classification (Hybrid) for Airline domain	87.39%	87.81%	**87.60%**

The experiments for hierarchical classification are conducted in 6 steps, i.e., Exp1.1 to Exp 1.6 (as mentioned in Table I) with the same dataset. Firstly, the sentiment classification on flat classification using NBC and SVM. Secondly, hierarchical classification with a rule-based approach using TextBlob. Then, performed the experiment on hierarchical classification using the proposed hybrid approach using TextBlob, Vader, Flair for subjectivity classification and NBC and SVM for sentiment classification. The findings clearly show that there is a consistent increasing score for f-measure from flat classification to rule-based hierarchical classification to hybrid hierarchical classification. The same dataset achieved NBC of 84.39% and SVM of 87.60%, while classification was performed with a hybrid approach on hierarchical classification. In contrast, while in flat classification, the subjectivity texts are not detected, and the mixture of subjective and objective texts influenced the overall accuracy of the result (which solely relies on the manual annotation of the classification under machine learning approach). In rule-based approach, subjectivity detection and negation detection are performed, which leads to a slight increase of the accuracy. However, there is still pitfall on the detection of sarcasm detection, which has been handled with hybrid approach and the accuracy visibly increased the sentiment analysis on subjective opinionated dataset. Since the result of hierarchical classification with SVM is promising, another evaluation measure is conducted with ROC and the achieved result shown as AUC = 0.907 for the ROC as shown in Figure 6, which is in an acceptable range as it distinguishes positive class values from negative class values.

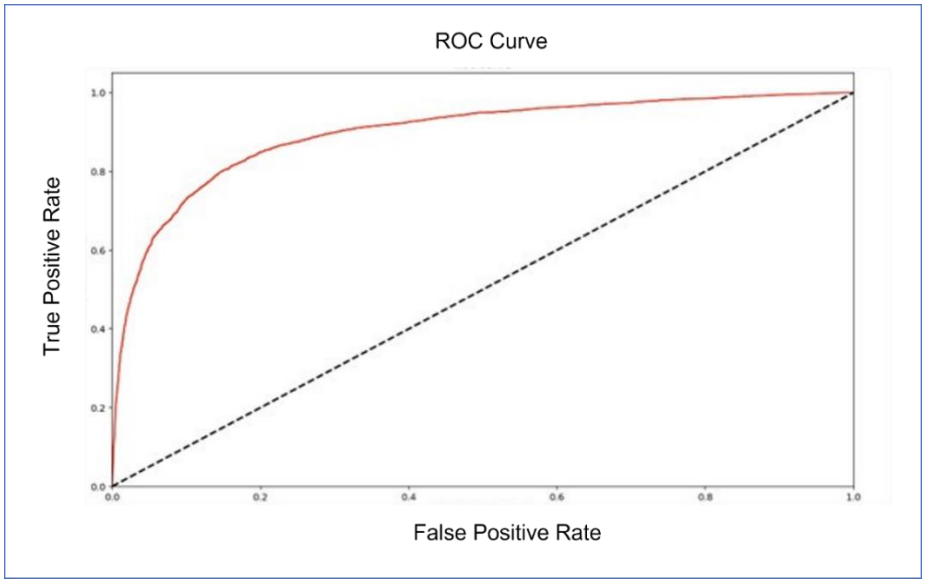

Fig 6: ROC on the hierarchical classification approach for Airline Domain

For topic modelling for the airline domain, 20 topics play the closest coherence score (as 0.539), and for the aspect categorization, there are 15 aspect categories extracted from the dataset as Flight Schedule, Flight Route, Baggage Handling, Airport Impression, Customer Service, Complaint Handling, Fare and Deals, Flight Experience, Inflight Meals, Comfort and Space, Inflight Amenities, Ground Service, Inflight Entertainment, Booking and e-information, and finally Inflight Crew Service. Therefore, it seems to have hidden aspects in the dataset which was not declared or discussed in state-of-art related papers [11][19][28] because the aspect categories from the state-of-art papers are adapted from the online review portals as in Figure 7.

Fig 7: The declared aspect categories from the Skytrax airline review portal (Skytrax, 2021)

Besides, In the Travelers Choice awards for 2018, TripAdvisor revealed the Top airlines across the world, according to an accumulation of the best online reviews. Based on this information, an article described the Top 5 Features That Make a Great Airline. In the article, customer service is rated as the number 1 feature that makes a great airline,

followed by comfort and space, great price and great deals, inflight meals and entertainment [33]. The customer service is different from the ground service and inflight crew service. In the current existing review criteria, this particular criterion is not explicitly mentioned. In contrast, the airline review dataset consists of the reviews that specifically highlighted the customer service issues such as the response to the inquires etc. Moreover, the criteria for full-service carrier product rating in the airline review rating portal mentioned that inflight amenities (e.g., bed, toilet, washroom, blanket, and pillows, etc.) plays a role in a user review for the airlines [34].

Therefore, the outcome of the experiment unhidden the aspects which was not revealed in previous studies while applying the manual aspect categorization technique. With the proposed solution, the result of the sentiment analysis highlighted the interesting business insights to improve airline services. As illustrated in Figure 8 for aspects versus review sentiment data visualization, “Flight Schedule” and “Booking and E-information” are the two aspects which received higher negative reviews from customers. These aspects are denoted as important aspects, especially during COVID-19 pandemic as most of the bookings are processed online and the tendency to modify the flight schedules due to travel restrictions is elevated. Thus, the aforementioned improvements will provide competitive advantage for the airline business. Additionally, with the alignment with real-world business articles, the result of the sentiment analysis from this study concludes as airlines should review the services in "Flight Schedule" and "Booking and E-information" to pay attention to business optimization to increase customer satisfaction.

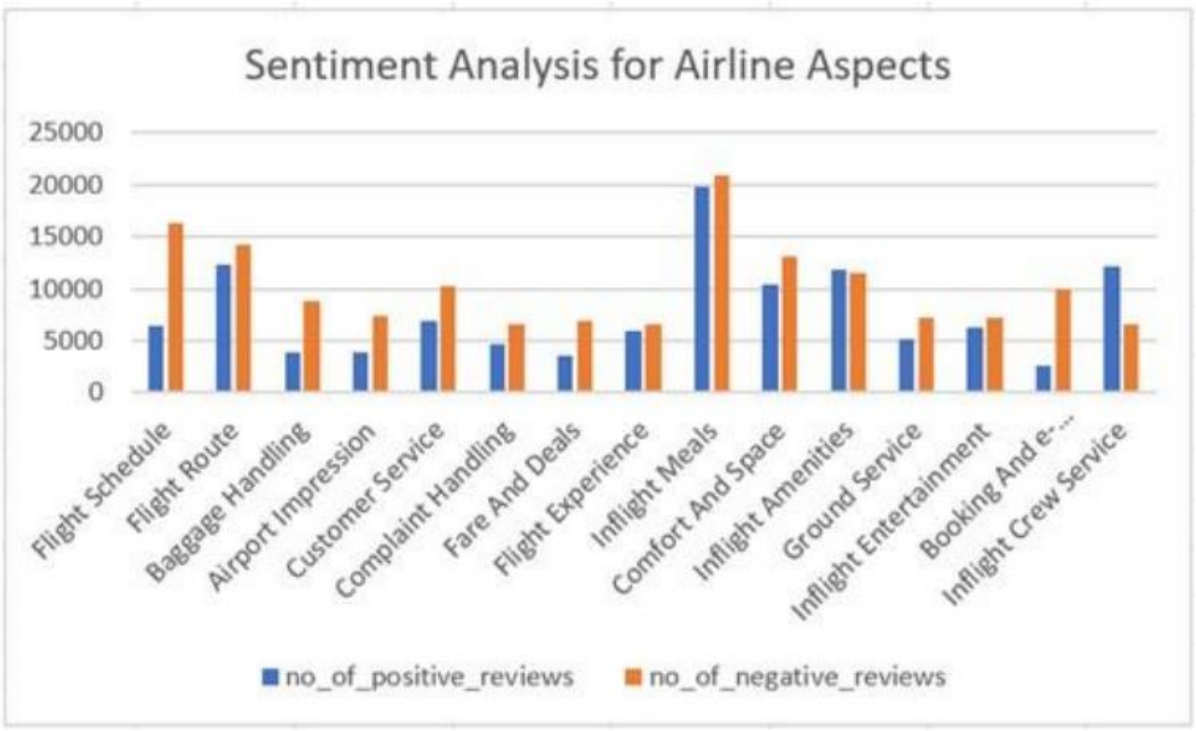

Fig 8: Sentiment analysis for airline aspects

With the satisfying result from airline domain, the baseline experiment is repeated for hotel domain to validate the applicability of the technique in another domain. Since the baseline is the comparison of Flat and Hierarchical classification with Hybrid approach, another set of experiment is focused on flat and hybrid technique with NBC and SVM respectively. The outcome of the hierarchical classification experiment for the hotel domain are described in Table 2.

Table 2: The result of Hierarchical classification experiment for Hotel Domain

Experiments	Precision	Recall	F-measure
Exp3.1: NBC on Flat Classification for Hotel Domain	61.89%	67.53%	**59.74%**
Exp3.2: SVM on Flat Classification for Hotel Domain	63.95%	60.37%	**61.72%**
Exp3.3: NBC on Hierarchical Classification (Rule-based) for Hotel domain	63.52%	67.13%	**65.27%**
Exp3.4: SVM on Hierarchical Classification (Rule-based) for Hotel domain	68.31%	65.21%	**66.72%**
Exp3.3: NBC on Hierarchical Classification (Hybrid) for Hotel Domain	72.15%	80.23%	**75.98%**
Exp3.4: SVM on Hierarchical Classification (Hybrid) for Hotel Domain	73.39%	79.02%	**76.10%**

Similar to the airline domain, the result of experiments for the hotel domain shown that the hierarchical classification using the Hybrid approach on the same dataset outperforms the flat classification approach. The weighted harmonic mean of the experiment's precision and recall draws the conclusion that a linear classifier (SVM) performs better than a probabilistic classifier (NBC). In this 2nd round of experiment, the hybrid approach is proven to handle the sarcasm, negations, word ambiguity and multipolarity issues in sentiment analysis. This evaluation represented that consideration of subjectivity classification with rule-based and machine learning approach can significantly increase the accuracy of a model. This finding is also aligned with the findings of the baseline paper mentioned in the related work [28] where the author stated that SVM performed better than NBC on the experiment of sentiment classification on customer reviews data of Soekarno-Hatta Airport. Similar to the Airline domain experiment, with the promising result from hierarchical classification with SVM, ROC is conducted additionally, and result of AUC is achieved as 0.897 as shown in Figure 9, which validated the reliable accuracy of the proposed model as in different performance measure.

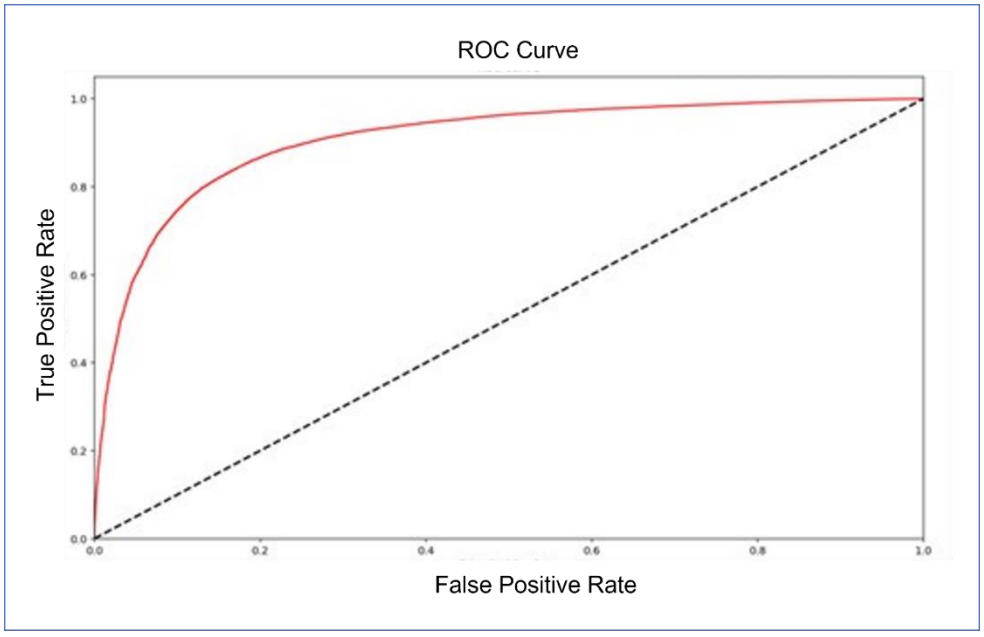

Fig 9: ROC on the hierarchical classification approach for Hotel Domain

Similarly, for the topic modelling, the experiment is conducted on the opinionated dataset. For the hotel domain, the optimal number of topics for LDA is 98 topics with a coherence score of 0.519, and for the aspect categorization, the result of the aspect categories returned a total of seven categories which are Hotel Amenities, Room Amenities, Services, Décor, Meals, Environment & Location and Price. The extracted aspect categories found to be relevant with the real-world reviews mentioned in [35], which stated that there is an aspect like Décor, that plays the role of customer satisfaction in hotel service. This sort of aspect is not revealed or discussed in recent state-of-art papers due to the aspect categorization limitation in accordance with the online review portals [36]. Besides, current popular hotel rating portals such as TripAdvisor and Agoda also use limited rating criteria (e.g., Décor, Environment and location are not explicitly defined) as shown in Figure 10 and Figure 11.

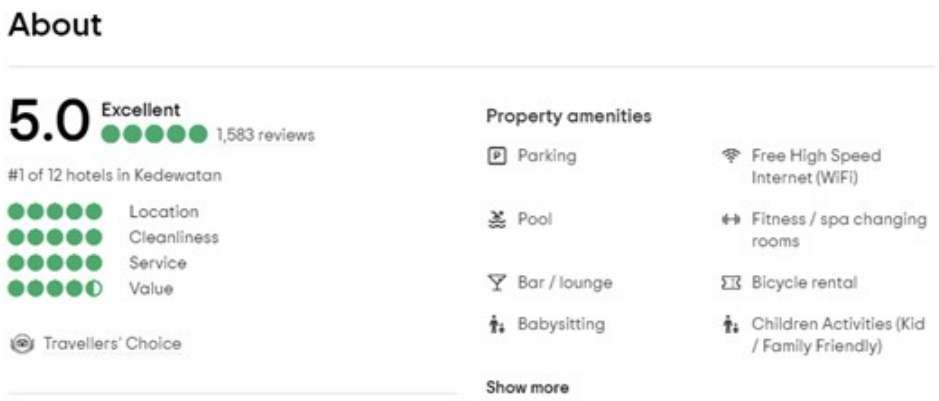

Fig 10: Rating for hotels in TripAdvisor (TripAdvisor, 2021)

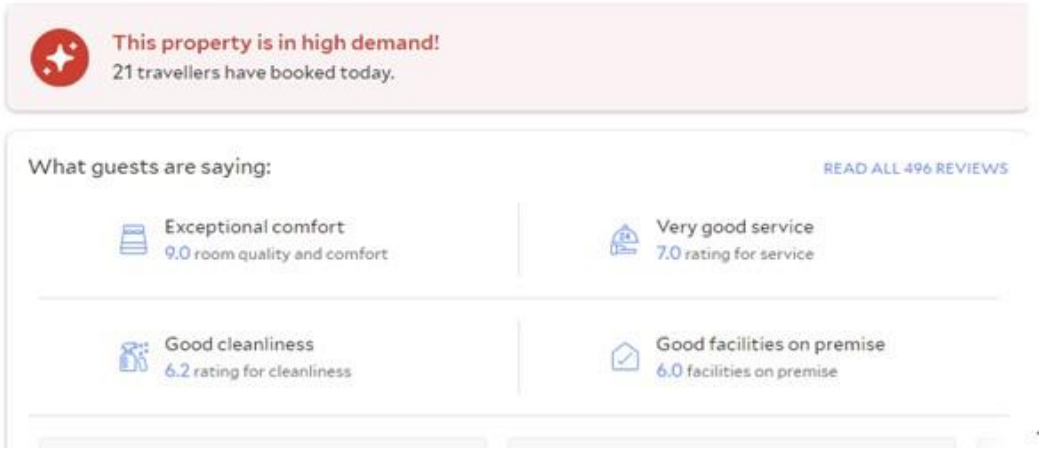

Fig 11: Rating for hotels in Agoda (Agoda, 2021)

Reference to the real-world business articles, the sentiment analysis concludes as Hotel should review the services in "Service" and "Meals" to pay attention to business optimization to increase customer satisfaction, as the number of negative reviews for “Service” and “Meals” aspects are relatively higher than others and indicate that these aspects carry higher negative sentiment from customers. Improving these aspects will anticipate customer dissatisfaction on hotel operation and promote business optimization. The visualization of aspects versus review sentiment is illustrated in Figure 12.

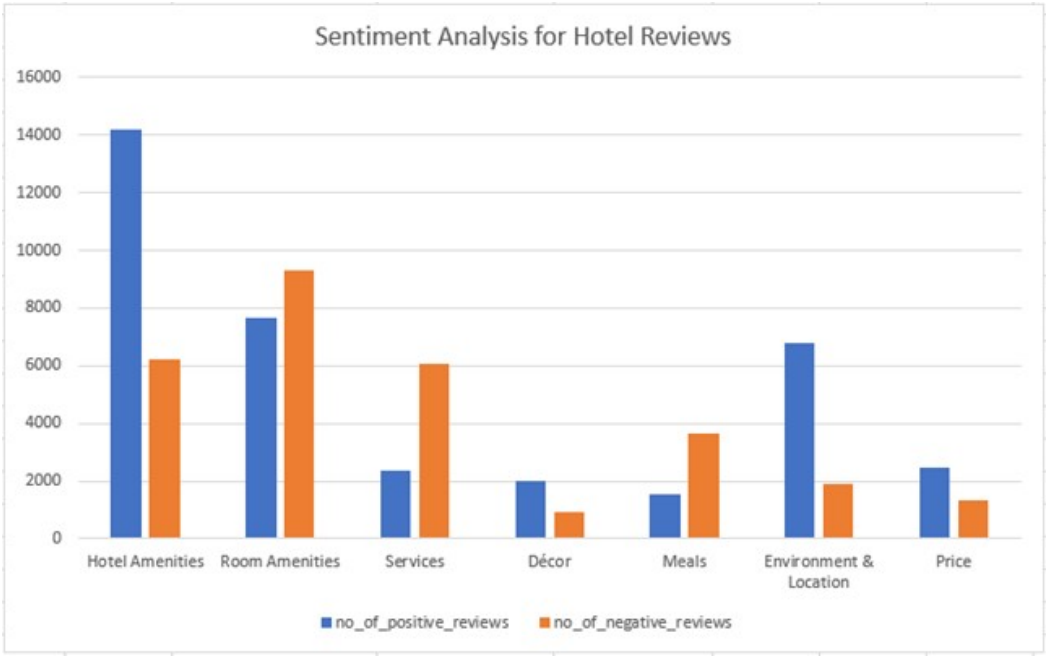

Fig 12: Sentiment analysis for hotel aspects

5.0 DISCUSSION

There is no previous work on the hierarchical classification with Hybrid approach. Therefore, phase by phase experiment is performed on three different scenarios on same dataset for benchmarking and providing implication to the performance. Experiments had performed on Flat classification (first scenario), hierarchical classification with rule-based (second scenario), and hierarchical classification with Hybrid (third scenario). Hypothesis is preserved as hierarchical classification with hybrid produces the best accuracy result. Hence, for Airline domain, F-score is achieved as 58.98% in first scenario, 69.52% in second scenario and 84.39% in third scenario. This shows significant increase in f-score value from flat classification (first scenario) to classification with hybrid (third scenario as proposed approach). Similarly, to hotel domain, f-score increased from 59.74% to 75.98% compared as first scenario and third scenario. Thus, based on the achieved results, this paper accepted the aforementioned hypothesis as F-score of hierarchical classification with hybrid approach outperforms the f-score of standalone flat classification and hierarchical classification using rule-based approach.

Additionally, same dataset is used to repeat the experiments of two baseline papers using the existing techniques mentioned by the authors, which used sentiment classification on NBC with subjectivity classification using only a rule-based approach [15] and sentiment classification with flat classification using SVM [28]. Comparison of the experimental result from this research and the baseline papers are discussed in Table 3.

Table 3: The comparison of Hierarchical classification experiment with baseline papers

NBC with Hierarchical Classification	F-measure
Proposed Technique (Airline Reviews)	**84.39%**
Proposed Technique (Hotel Reviews)	**75.98%**
Existing Technique by Lubis *et al*. (2017) (Airline Reviews)	**69.52%**
Existing Technique by Dhini and Kusumaningrum (2018) (Hotel Reviews)	**61.72%**

Comparing the F-measure across the models, the proposed model seems to outweigh the other two papers. After optimizing with the hybrid approach, the result of the F-measure increased significantly. The paper of [15] stated that the limitation in this paper is NBC does not get the necessary information about the minority class to build an accurate prediction. Therefore, there is an imbalance distribution in precision and recall score, also the low weighted value of F-measure. Thus, the authors suggested using for sampling method to modify imbalanced data into balanced

distribution using some mechanism [15]. In the proposed model, stratified sampling, SMOTE and cross-validation is used to tackle this issue. Henceforth, the overall result of the proposed model looks promising.

Moreover, the observations from this experiment align with the related research which stated SVM and Vader performs better for online user reviews [26] and concludes that subjectivity detection with hybrid approach returns better result than rule-based approach. Henceforth, the performance of this experiment is acceptable as this bypass the selected baseline paper with the higher F-measure of the classifier. All in all, the findings from the above experiments prove that hierarchical classification in sentiment analysis technique on online user reviews can improve accuracy by efficiently eliminating the objective comments (non-opinionated or factual texts) from the user review dataset.

On the other hand, identifying the aspects from the reviews itself after creating the opinionated dataset seems to provide the aspects that are not highlighted before in existing studies using predefined aspect categories. Also, automation of topic-aspect mapping and topic cluster labelling eliminate labor intensiveness in manual categorization/mapping tasks in current studies and outperform the result of dependency parsing. Semantic similarity clustering with GloVe and K-means added benefits to the automation of synonym clustering. The learning from topic modelling is that the cleaning process is a major part of LDA. Keeping only nouns and verbs (POS Tagging), removing templates from texts, testing different cleaning methods iteratively will improve the topics. Therefore, there is an advantage of performing topic modelling on the hierarchical classification proposed in this study. Additionally, the comparison of the experimental result for topic modelling from this research and the baseline papers are discussed in Table 4, where the existing paper used opinionated extraction before LDA, however, topic-aspect mapping (as automatic aspect categorization) was done manually, and performance measure is solely evaluated with perplexity. In contrast, the proposed model introduced automatic topic-aspect mapping and measured the performance using perplexity as well as coherence.

Table 4: The comparison of Topic Modelling experiments with baseline papers

Topic Modelling	**Opinionated extraction**	**Automatic aspect categorization**	**Performance Metric**	**Perplexity**	**Coherence**
Proposed Model (Airline Review)	Yes	Yes	Perplexity, Coherence	-5.439	**0.539**
Proposed Model (Hotel Review)	Yes	Yes	Coherence	-	**0.519**
Lubis *et al*., 2019	Yes	-	Perplexity	Lowest perplexity when the topic number is 100.	-

Hence, the approach and result of topic modelling are compared with the baseline papers and satisfied the future work of [16], which is to include the coherence evaluation for the performance. The result clearly shows that the proposed model is in the range of acceptable score, yet, can be looked into the hyper tuning or optimization for better results. Moreover, authors of [11] suggested performing automatic aspect extraction and topic cluster labelling using a designed algorithm to improve the efficiency of topic-aspect mapping and overall aspect-based sentiment analysis. This has been addressed in this study by proposing automatic aspect categorization to extract aspects from reviews itself and build aspect category on it to ensure all the dimensions in the reviews are captured and included in the analysis. Also, topic modelling to unhide topics from reviews and mapped to various aspect category automatically as topic cluster labelling.

From the perspective of practical and society implications, the objective of this research is to sustain the economic recovery through data-driven and digital transformation approach during the COVID-19 pandemic. The COVID-19 pandemic has caused a huge hit to most major economies and businesses are thriving to figure out different strategies that can drive business excellence during the pandemic. With the limitation of physical interactions, business sectors rely solely on social listening thru online reviews to understand better their customers and services. Therefore, advancement in Artificial intelligence or machine learning technologies in sentiment analysis can better facilitate the process of social listening. Thus, this research introduces advancement in sentiment attribution analysis technique with hierarchical classification and automatic aspect categorization, to assist in identifying the factors that influence the business so that business can craft a new direction for marketing and operation strategy during this pandemic and

sustain the economic recovery throughout this pandemic. The contribution to the body of knowledge from this research can be used in the practice of economic impact for diligent market research, as it is imperative to get a clear picture of what strategies will succeed and what won't during this pandemic – regardless of a new product launch, new service, product positioning, or even for a new target market. Moreover, this paper bridges the gap between the theory of improving overall sentiment attribution analysis (with unhiding valuable aspects for business operations and eliminating non-opinionated reviews from opinion summary) and the practical implication of applying the outcome of this sentiment attribution analysis in dedicated business sectors to improve business service provisioning during the pandemic.

6.0 CONCLUSION

In brief, although sentiment analysis is well established for multiple domains, little work has been carried out to study the different options of hierarchical classification, and previous works have not comprehensively considered the hybrid approach in hierarchical classification. Similarly, there is no comprehensive work dedicated to automatic aspect categorization from the review itself and automatic topic cluster labelling for improved topic-aspect mapping and attribution analysis. Henceforth, this research proposed a hybrid approach for hierarchical classification, which use of rule-based and machine learning techniques to build opinionated reviews prior to sentiment classification by eliminating the irrelevant texts from the training of the model and improve the accuracy of the sentiment result. The current findings from this research declared that hierarchical classification with a hybrid approach outperforms the flat classification in sentiment analysis with the f-score of average of 17% and 14% improvement, respectively in Airline and Hotel domain. As the end results are promising, this research continues with the dimensions of automatic aspects extraction and algorithmic techniques to map the topics into extracted aspects, which certainly improves the sentiment analysis result and provide reliable input for business insights in the hotel domain and airline domain. The overall result for each proposed technique has shown promising performance results and is reflective of the current business articles. Additionally, this research satisfied the future work suggested in the three baseline papers [15][16][11].

However, as for the limitation, currently, the proposed methodology is only tested on the Airline Domain and Hotel Domain online user review datasets and the applicability of the solution on any other cross-domains are not covered. As future work, the improvement in aspect extraction with different clustering methods can be achieved. Studying the performance measures with different classification algorithms (such as Random Forest, C4.5, etc.) and emphasizing the optimization of ABSA analysis is another potential future work.

ACKNOWLEDGEMENT
This research was supported by the Impact-Oriented Interdisciplinary Research Grant Programme (IIRG) Universiti Malaya (IIRG005B-2020SAH) and Faculty Research Grant Universiti Malaya (GPF098C-2020).

REFERENCES

[1] Liu, Z., Lei, S., Guo, Y., & Zhou, Z. (2020), "The interaction effect of online review language style and product type on consumers' purchase intentions", Palgrave Communications, Vol. 6, pp. 11. https://doi.org/10.1057/s41599-020-0387-6

[2] Murphy, R.. (2020), "Local Consumer Review Survey: How Customer Reviews Affect Behavior", Available at: https://www.brightlocal.com/research/local-consumer-review-survey/?SSAID=389818&SSCID=31k5_6cxgn (accessed 7 March 2021)

[3] Shi, WH., Zhang, Q., and Cai JL. (2018), "The impact of contradictory online reviews on ambivalent attitude and purchase intention", Management Review, Vol. 30 No. 7, pp.77–88.

[4] Ghasemaghaei, M., Eslami, PS., Deal K. and Hassanein, K. (2018), "Reviews' length and sentiment as correlates of online reviews' ratings", Internet Res, Vol. 28 No. 3, pp.544–563.

[5] Saleem, M., Zahra, S. and Yaseen, A. (2017), "Impact of service quality and trust on repurchase intentions – the case of Pakistan airline industry", *Asia Pacific Journal of Marketing and Logistics*, Vol. 29 No.5, pp.1136-1159. https://doi.org/10.1108/APJML-10-2016-0192

[6] Albers, S. and Rundshagen, V. (2020), "European airlines′ strategic responses to the COVID-19 Pandemic (January-May, 2020)", *Journal of Air Transport Management*, Vol. 87, pp.101863. https://doi.org/10.1016/j.jairtraman.2020.101863

[7] Ban, H. and Kim, H. (2019), "Understanding Customer Experience and Satisfaction through Airline Passengers' Online Review". *Sustainability*, Vol. 11 No. 15, pp.4066. https://doi.org/10.3390/su11154066

[8] Shahnawaz and Astya, P. (2017), "Sentiment analysis: Approaches and open issues", in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, IEEE, New York, NY pp.154-158. https://doi.org/10.1109/CCAA.2017.8229791

[9] Rane, A. and Kumar, A. (2018), "Sentiment Classification System of Twitter Data for US Airline Service Analysis", in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, IEEE, New York, NY, pp.769-773. https://doi.org/10.1109/COMPSAC.2018.00114

[10] Pascual, F. (2019), "A Comprehensive Guide to Aspect-based Sentiment Analysis", available at: https://monkeylearn.com/blog/aspect-based-sentiment-analysis/ (accessed 7 March 2021)

[11] Anoop, V. and Asharaf, S. (2018), "Aspect-Oriented Sentiment Analysis: A Topic Modeling-Powered Approach", *Journal of Intelligent Systems*, Vol. 29 No. 1, pp.1166-1178. https://doi.org/10.1515/jisys-2018-0299

[12] Tao, W., Zhang, Q., Zhang, M. and Li, Y. (2019), "Mining Pain Points from Hotel Online Comments Based on Sentiment Analysis", in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, IEEE, New York, NY, pp.1672-1677. https://doi.org/10.1109/ITAIC.2019.8785893

[13] Samonte, M., Garcia, J., Lucero, V. and Santos, S. (2017), "Sentiment and opinion analysis on Twitter about local airlines", in Othman, J.B & Gang, F. (Ed.s), *2017 Proceedings of the 3rd International Conference on Communication and Information Processing - ICCIP '17*, Association for Computing Machinery (ACM), New York, NY, pp.415-422. https://doi.org/10.1145/3162957.3163029

[14] Sharma, N., Rahamatkar, S., & Sharma, S. (2019), "Classification of Airline Tweet Using Naïve-Bayes Classifier for Sentiment Analysis", in 2019 International Conference On Information Technology (ICIT), IEEE, New York, NY, pp. 70-75. https://doi.org/10.1109/ICIT48102.2019.00019

[15] Lubis, F., Rosmansyah, Y. and Supangkat, S.(2017), "Improving course review helpfulness Prediction through sentiment analysis", in Setti, G. (Ed), *2017 International Conference on ICT For Smart Society (ICISS)*, IEEE, New York, NY, pp.1-5. https://doi.org/10.1109/ICTSS.2017.8288877

[16] Lubis, F., Rosmansyah, Y. and Supangkat, S. (2019), "Topic Discovery of Online Course Reviews Using LDA with Leveraging Reviews Helpfulness", *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 9 No. 1, pp.426. https://doi.org/10.11591/ijece.v9i1.pp426-438

[17] Shahul, E. (2021), "Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch", available at: https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair (accessed 7 March 2021)

[18] Mahoto, N., Gul Khan, S. and Ruk, S. (2018), "A Lexicon-based Method to Determine Subjectivity Of Unstructured Data", in Setti, G. (Ed), *2018 5th International Multi-Topic ICT Conference (IMTIC)*, IEEE, New York, NY, pp.1-7. https://doi.org/10.1109/IMTIC.2018.8467271

[19] Lacic, E., Kowald, D. and Lex, E. (2016), "High Enough? explaining and predicting traveler Satisfaction using airline reviews", in Bluestein, J., & Eerder, E. (Ed.s), *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, Association for Computing Machinery (ACM), New York, NY, pp.249-254. https://doi.org/10.1145/2914586.2914629

[20] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar S. (2014), "Semeval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pp. 27–35. https://dx.doi.org/10.18653/v1/S16-1002

[21] Ganu,G., Elhadad, N., and Marian, A. (2009), "Beyond the stars: Improving rating predictions using review text content", in *2009 Twelfth International Workshop on the Web and Databases (WebDB* 2009), WebDB, Providence, Rhode Island, NY, pp.1-6.

[22] Sasmita, D., Wicaksono, A., Louvan, S. and Adriani, M. (2017), "Unsupervised aspect-based sentiment analysis on Indonesian restaurant reviews", in *2017 International Conference on Asian Language Processing (IALP)*, IEEE, New York, NY, pp.383-386. https://doi.org/10.1109/IALP.2017.8300623

[23] Ching, M., & de Dios Bulos, R. (2019), "Improving Restaurants' Business Performance Using Yelp Data Sets through Sentiment Analysis", in Proceedings of the 2019 3rd International Conference on E-Commerce, E-Business And E-Government - ICEEG 2019. Association for Computing Machinery (ACM), New York, NY, pp.62-67. https://doi.org/10.1145/3340017.3340018

[24] Danisman, E. (2020), *Skytrax Airline Reviews*, Kaggle.com, available at: https://www.kaggle.com/efehandanisman/skytrax-airline-reviews (accessed 7 March 2021)

[25] Shaver, N. (2018), *Hotel Reviews*, Kaggle.com, available at: https://www.kaggle.com/datafiniti/hotel-reviews?select=7282_1.csv (accessed 7 March 2021)

[26] Rao, P. (2019), "Fine-grained Sentiment Analysis in Python (Part 1)", available at: https://towardsdatascience.com/fine-grained-sentiment-analysis-in-python-part-1-2697bb111ed4 (accessed 7 March 2021).

[27] Hu, M. and Liu, B. (2004), "Mining and summarizing customer reviews", in Kohavi, R., & Kim, W. (Ed.s), *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, Association for Computing Machinery (ACM), New York, NY, pp.168-177. https://doi.org/10.1145/1014052.1014073

[28] Dhini, A. and Kusumaningrum, D.A. (2018), "Sentiment Analysis of Airport Customer Reviews", in Setti, G. (Ed), *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, New York, NY, pp.502-506. https://doi.org/10.1109/IEEM.2018.8607335

[29] Repustate Inc. (2021). What makes Aspect-based sentiment analysis better?. Retrieved 9 August 2021, from https://www.repustate.com/blog/choosing-between-topic-and-aspect-sentiment-analysis/

[30] Mouthami, K.; Devi, K. N.; Bhaskaran, V. M. (2013). "Sentiment analysis and classification based on textual reviews", in 2013 IEEE International Conference on Information Communication and Embedded Systems (ICICES 2013) - Chennai (2013.2.21-2013.2.22)], pp. 271–276. https://doi.org/10.1109/ICICES.2013.6508366

[31] Bhandari, A. (2020), "AUC-ROC Curve in Machine Learning Clearly Explained – Analytics Vidhya", available at: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/ (accessed 10 April 2021)

[32] Syed, S. and Spruit, M. (2017), "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation", in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, New York, NY, pp.165-174. https://doi.org/10.1109/DSAA.2017.61

[33] Park, Sangwon; Lee, Jin-Soo; Nicolau, Juan L. (2020). Understanding the dynamics of the quality of airline service attributes: Satisfiers and dissatisfiers. Tourism Management, Vol.81, pp.104163. https://doi.org/10.1016/j.tourman.2020.104163

[34] Airline Ratings (2021). "Full Service Carrier Product Rating Criteria - Airline Ratings", available at: https://www.airlineratings.com/full-service-carrier-product-rating-criteria/ (accessed 7 March 2021)

[35] Doğan, S., Basaran, M.A. and Kantarci, K. (2020), "Determination of attributes affecting price-performance using fuzzy rule-based systems: online ratings of hotels by travel 2.0 users", Journal of Hospitality and Tourism Technology, Vol. 11 No. 2, pp. 291-311. https://doi.org/10.1108/JHTT-07-2018-0067

[36] Ye, P. and Yu, B. (2018), "Customer Satisfaction Attribution Analysis of Hotel Online Reviews Based on Qualitative Research Methods", in *Proceedings of the 2nd International Conference on E-Education, E-Business and E-Technology - ICEBT 2018*, Association for Computing Machinery (ACM), New York, NY, pp.93-98. https://doi.org/10.1145/3241748.3241758

[37] Mahoto, N., Gul Khan, S. and Ruk, S., 2018. A Lexicon-based Method to Determine Subjectivity of Unstructured Data. 2018 5th International Multi-Topic ICT Conference (IMTIC), pp.1-7. https://doi.org/10.1109/IMTIC.2018.8467271