sciendo

ISMSME International Society for Manufacturing, Service and Management Engineering

received: 2 January 2023
accepted: 1 October 2023

pages: 31-47



# Supervised multilabel classification techniques for categorising customer requirements during the conceptual phase in the new product development

Sutrilastyo
Ririn Diar Astanti

ABSTRACT
The research aims to provide the decision-maker with a framework for determining customer requirements during product development. The proposed framework is based on sentiment analysis and supervised multilabel classification techniques. Therefore, the proposed technique can categorise customer reviews based on the "product design criteria" label and the "sentiment of the review" label. To achieve the research goal, the research presented in this article uses the existing product development framework presented in the literature. The modification is conducted especially in the conceptual stage of product development, in which the voice of the customer or a customer review is obtained from the scraping, and a multilabel classification technique is performed to categorise customer reviews. The proposed framework is tested by using the set data on women's clothing reviews from an e-commerce site downloaded from www.kaggle.com based on data by Agarap (2018). The result shows that the proposed framework can categorise customer reviews. The research presented in this paper has contributed by proposing a technique based on sentiment analysis and multilabel classification that can be used to categorise customers during product development. The research presented in this paper answers one of the concerns in the categorisation of needs raised by Shabestari et al. (2019), namely, the unclear rules or main attributes of a requirement that make these needs fall into certain categories. Categorising customer requirements allows decision-makers to determine the direction of product development to meet customer needs.

KEY WORDS
**voice of customer, customer needs, conceptual design, product development process, customer's requirement categorisation, sentiment analysis, supervised multilabel classification technique**

**10.2478/emj-2024-0003**

**Ririn Diar Astanti**

Faculty of Industrial Technology
Universitas Atma Jaya Yogyakarta
Jalan Babarsari Street 44
55281 Yogyakarta, Indonesia
ORCID 0000-0001-5234-0283

Corresponding author:
e-mail: ririn.astanti@uajy.ac.id

**Sutrilastyo**

Faculty of Industrial Technology
Universitas Atma Jaya Yogyakarta
Jalan Babarsari Street 44
55281 Yogyakarta, Indonesia
e-mail: 195603118@students.uajy.ac.id

## Introduction

New product development (NPD) can be described as a formalised planning process that encompasses the steps from initial idea generation to product launch in the market (Kim et al., 2016). New product development is critical to supporting business growth (Cooper, 2001; Ulrich & Eppinger, 2015). The popular new product development (NPD) stage consists of several stages, such as the initial product concept or idea that is evaluated, developed, tested, and launched on the market (Booz et al., 1982;

Sutrilastyo, & Astanti, R. D. (2024). Supervised multilabel classification techniques for categorising customer requirements during the conceptual phase in the new product development. *Engineering Management in Production and Services, 16*(1), 31-47. doi: 10.2478/emj-2024-0003

Büyüközkan & Arsenyan, 2012; Kaulio, 1998; Shabestari et al., 2019; Ulrich & Eppinger, 2015). During the conceptual phase, the ideas are developed into product specifications or dimensions, such as product features, design, and durability (Di Benedetto, 1999; Schulze & Hoegl, 2006). The conceptual design stage is crucial in the product development process as it identifies customer needs. Later, if considered during product development, these needs can provide a competitive advantage for the company (Bhuiyan, 2011).

One example of consumer needs that should be considered during the conceptual phase is related to product design. According to Kreuzbauer and Malter (2005), attractive product designs impact the image and positive brand for customers. The product failure risk can be reduced by considering the voice of the customer (VoC) in the product design and development process (Nazari-Shirkouhi & Keramati, 2017). VoC and value derived from a product are also essential quality aspects (Kapucugil Ikiz & Özdağoğlu, 2015; Mulay & Khanna, 2017). VoC contains the expectations and needs that customers feel for a product or service. Obtaining VoC for a company is a crucial activity (Aguwa et al., 2017).

Several methods for obtaining VoC are surveys, interviews, and focus group discussions (Kapucugil Ikiz & Özdağoğlu, 2015; Šperková, 2019). However, these methods face several obstacles in many situations, such as cost efficiency, time, and space constraints. To overcome these limitations, several previous studies (Devi et al., 2016; Jeong & Yoon, 2016; Misopoulos et al., 2014) collected VoC through media on the Internet, such as product reviews, forum discussions, and social media.

One of the methods to analyse VoC is text mining. According to Khedr et al. (2017), text mining is a method for extracting knowledge from text. Pinquié et al. (2018) used text mining to obtain the quality requirements of documents. Text mining was used by Lee and Bradlow (2011) to extract information from customer reviews and then cluster it into groups. Kapucugil Ikiz and Özdağoğlu (2015) suggested that text mining techniques help speed up the VoC collection process.

One of the text mining methods used to identify VoC is sentiment analysis. Sentiment analysis is one of the text mining methods used to determine the contextual polarity of an article, whether negative, neutral, or positive (Shukri et al., 2015). The use of sentiment analysis, as described by Jeong and Yoon (2016), can identify features that can be further developed on smartphones. Huang et al. (2013) used a multi-task and multilabel classification approach in conducting a sentiment analysis and topic classification on the Twitter platform. The study found that sentiment labels helped in the process of accurately distinguishing topics. Other researchers stated that sentiment analysis is used to identify the weaknesses of products reviewed by customers (Zhang et al., 2012) and to measure customer satisfaction (Kang & Park, 2014).

Shabestari et al. (2019) summarised studies using Machine Learning (ML) and text-mining techniques in the initial stages of product development. Most of the literature uses one of the unsupervised learning techniques in the product requirements categorisation stage using ML. Only one study used supervised learning techniques. Research using the supervised learning method in the needs categorisation stage has not been widely carried out. Meanwhile, according to Edwards et al. (2021) and the research results by Abad et al. (2017), classification using supervised learning techniques tends to be more accurate when compared to unsupervised learning techniques. In addition, Shabestari et al. (2019) also stated that there were no studies comparing the performance of one classification/clustering method with another.

Although sentiment analysis can be used by designers to quickly obtain customer preference data, both qualitatively and quantitatively, the insights obtained from these sentiments are still limited (Ireland & Liu, 2018). This is because the characteristics of certain features of a product cannot be known. Therefore, to fill this gap, adding a parameter that can show the characteristics of a product's features, e.g., by using product criteria, is considered in the proposed framework presented in this paper. The research proposed in this paper tries to fill the existing gaps by categorising customer requirements based on the "product design criteria" label and "sentiment of the review" using a supervised multilabel classification technique.

## 1. Literature review

This chapter discusses previous research on Voice of Customer (VoC) and the use of Sentiment Analysis and Machine Learning (ML) during product development.

### 1.1. Product development process

Product development is a vital part of maintaining the competitiveness of any organisation. This process determines product specifications and production processes by considering market needs, technology availability, and the organisation's strategy (Kim et al., 2016). Ulrich and Eppinger (2015) defined product development as a series of activities that begin with identifying market opportunities and end with product production, sales, and delivery. They divided the product development process into several stages: planning, concept development, system-level design, detail design, testing and refinement, and production. These are divided into several smaller stages. For example, concept development is divided into stages of identifying customer needs, determining product specifications, generating product concepts, and conducting concept tests. Conceptual design deals with how needs are identified and the formulation of product design concepts. The prototype tests the proposed concept, and the product launch is related to the product commercialisation process.

Kaulio (1998) also divided the product development process into five phases: specification, concept development, detailed design, prototype, and final product. Büyüközkan and Arsenyan (2012) proposed similar stages in their research on collaborative product development. Pienaar et al. (2019) used a stage-gate model where product development consists of four "gates" and four stages, i.e., the exploration stage, where ideas are conceptualised; the assessment stage, where opportunities are defined; the research stage, where the pilot/first technology development is carried out, and the last stage is implementation preparation where the pilot technology commercialisation trial is made.

Various stages of new product development are suggested in the studies mentioned above. According to Carter (2015), no one best product development stage can outperform other product development stages. In general, product development stages can be grouped into three major groups: conceptual design, prototype, and product launch. The grouping of the stages proposed by previous researchers into three major groups of product development stages can be seen in Table 1.

According to Shabestari et al. (2019) and Kornish and Hutchison-Krupat (2017), the conceptual design consists of several stages: (1) requirement identification; (2) requirement categorisation; and (3) conceptual selection, especially the requirements categorisation during product development. During the requirements identification, the Voice of Customer (VoC) is needed.

VoC contains consumer expectations and needs for products, and these expectations help companies in the process of developing their products or services (Aguwa et al., 2017). VoC is a valuable resource for companies. The methods used to collect VoC can be in the form of surveys, interviews, focus group discussions, and other methods (Kapucugil Ikiz & Özdağoğlu, 2015; Šperková, 2019).

With the development of technology, the platforms for VoC are also increasing. Social media and e-commerce platforms, such as Facebook, Twitter, Reddit, YouTube, Amazon, and Tokopedia, are used to collect data in the form of user posts, reviews, and comments. The feedback that customers give voluntarily on these platforms is a key factor in the product development and design phase (Park et al., 2018). In addition, data taken from social media is used for various purposes, such as predicting the company's popularity level based on consumer reactions on social media (Park & Alenezi, 2018). Gupta and

Tab. 1. Product development stages

| Author | Product Development Stage | | |
|---|---|---|---|
| | Conceptual design | Prototype | Product launch |
| Ulrich & Eppinger (2015) | • Planning<br>• Concept development<br>• System-level design | • Detail design<br>• Testing and refinement | • Production ramp-ups |
| Kaulio (1998) | • Specification<br>• Concept development | • Detailed design<br>• Prototyping | • Final product |
| Büyüközkan & Arsenyan (2012) | • Conceptual design | • Product development<br>• Prototype | • Manufacturing<br>• Product launch |
| Pienaar et al. (2019) | • Explore<br>• Assess | • Research | • Prepare for implementation |

Sebastian (2018) looked at the performance of a product in the market based on ratings and reviews provided by users, and Zhou et al. (2016) compared consumer behaviour when shopping online.

After requirement identification, the next step in the conceptual stage of product development is requirement categorisation, which is processing the feedback from the requirement identification step. Large amounts of data are one of the obstacles in processing data obtained from online platforms, such as social media and e-commerce. Obtaining useful and meaningful information from the data is a challenge. Text mining helps bridge this process. Jeong and Yoon (2016) used text mining to get VoC and find development opportunities in smartphone products. Park et al. (2018) used sentiment analysis to find more important features in improving tire design.

### 1.2. Sentiment analysis

Sentiment analysis is also known by several other names, such as opinion mining, sentiment mining, and opinion extraction. It is a process of detecting and classifying the contextual polarity of the text (Micu et al., 2017). There are two kinds of sentiment analysis approaches: ML-based sentiment analysis and lexicon/dictionary-based sentiment analysis. ML-based sentiment analysis often involves a model assigning labels to data based on data used to train the model. Meanwhile, lexicon-based sentiment analysis uses pre-determined words, where each word is associated with a certain sentiment (Gonçalves et al., 2013).

Several methods often used to perform ML-based sentiment analysis are classification methods, such as Support Vector Machine (SVM) and Naïve Bayes (Kolchyna et al., 2015). Other methods, such as Maximum Entropy (ME), Logistic Regression, Apriori, and Random Forest, were also used in several other studies (Malviya et al., 2020; Samuel et al., 2020). Devi et al., 2016 used SVM to determine user perceptions of features from reviews of several laptop brands.

According to Ireland and Liu (2018), the insights obtained from the sentiment analysis are still limited. The result from sentiment analysis cannot identify the characteristics of a product's certain features. Therefore, to fill this gap, adding a parameter that can show the characteristics of a product's features, for example, by using product criteria, is considered in the proposed product development framework presented in this paper. Shabestari et al. (2019) summarised previous studies on using ML in the initial stages of product development. Among these studies, 51 use the ML method, both supervised and unsupervised learning methods. Twelve studies focus on the needs categorisation stage: 11 use the unsupervised learning method, and one — the supervised learning method. Aguwa et al. (2017) used a clustering technique to group keywords that play a key role in the quality of customer reviews. Fuzzy logic is then combined with the clustering model to capture the main essence of customer reviews so that the fulfilment of customer needs for products or services can be improved. Lee and Bradlow (2011) conducted text mining to get the pros and cons of customer reviews. The clustering process is carried out on the pros and cons attributes to divide the words into special categories. The process can reveal the position of the product compared to competitors or related products and highlight product attributes that stand out in the eyes of customers. Abad et al. (2017) classified customer needs into two categories: functional/functional requirements (FR) and non-functional/non-functional requirements (NFR). The research found that the classification of NFR into sub-categories, such as usability, availability, and performance, can be automatically improved by using ML. The proposed product development framework presented in this paper is adapted from Shabestari et al. (2019) and Kornish and Hutchison-Krupat (2017) by adding sentiment analysis considering the "product design criteria" label and "sentiment of the review" using supervised multilabel classification technique during the requirement categorisation process in the conceptual stage of the product development process.

## 2. Proposed framework

This research focuses on the conceptual design stage, especially the requirements categorisation process during product development. As mentioned in the previous section, the proposed product development framework presented in this paper is adapted from Shabestari et al. (2019) and Kornish and Hutchison-Krupat (2017). The modification has been made by adding sentiment analysis considering the "product design criteria" label and "sentiment of the review" using a supervised multilabel classification technique during the requirement categorisation

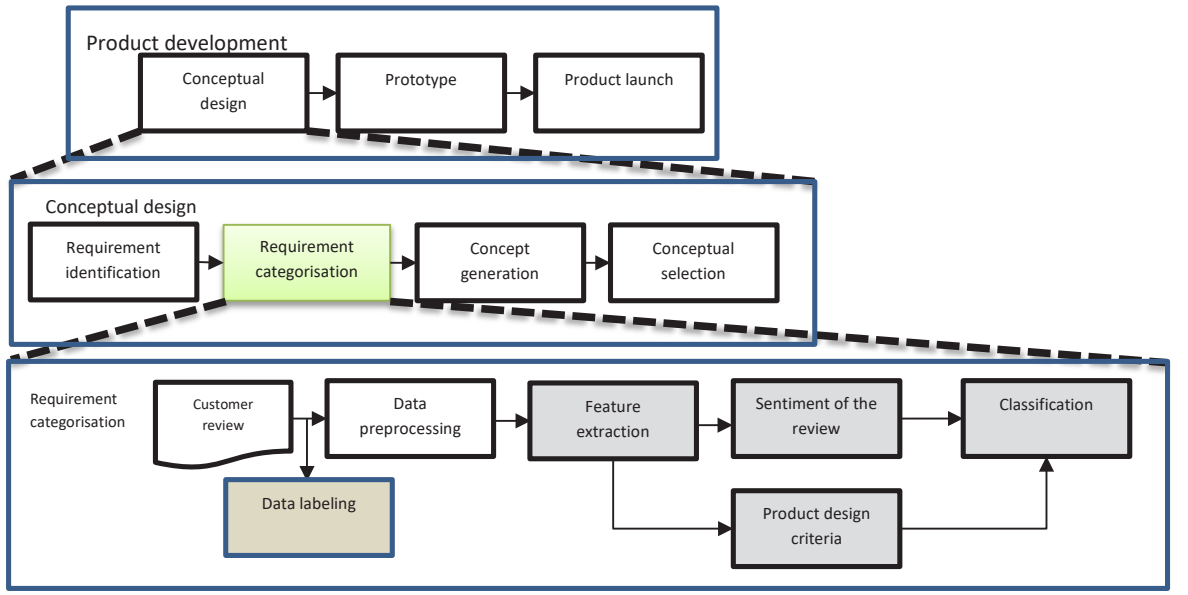



Fig. 1. Proposed framework for requirement categorisation in the product development process

process in the conceptual stage of the product development process.

The proposed framework is presented in Fig. 1. During the requirements categorisation process in the proposed framework, the categorisation process starts when the Voice of Customer (VoC) or a customer review is received from the identification process. The requirement categorisation step is performed next. The categorisation requirement steps consist of several activities: data labelling, data pre-processing, feature extraction, and classification. The difference between the framework proposed in this paper and the previous research is that during the requirement categorisation process:

- Customer review data is firstly labelled according to "product design criteria" and "sentiment of the review" before the pre-processing.
- Multilabel classification technique is based on "product design criteria" and "sentiment of the review". This approach has not been found in previous research. "Product design criteria" and "sentiment of the review" are the Feature Extraction results. The purpose of using multilabel classification in the need categorisation stage is to assist decision-makers in categorising customer needs based on one aspect of a product and customer perceptions of that aspect. Thus, this will assist in the stage of making product design concepts that can meet customer needs/wants. There are three classification techniques used to classify customer needs: Naïve Bayes, Logistics Regression, and Support Vector Machine. Among those three techniques, one has the highest accuracy and is selected as the classification technique in the proposed framework presented in this paper. The result from these classification techniques will be used as input for the next stage of the product development process, which is concept generation.

Details of each step in the proposed framework are presented in the following subsections.

### 2.1. Getting a customer review by web scraping techniques

As presented in the proposed framework in Fig. 1, especially during requirement categorisation, the step starts by getting customer product reviews using web scraping techniques. Web scraping is a process for automatically extracting and organising data from web pages (vanden Broucke & Baesens, 2018).

### 2.2. Data labelling

The obtained data is then labelled as "product design criteria" and "sentiment of the review". For example, women's clothing has four "product design criteria" classes consumers consider when deciding on a product.

### 2.3. Pre-processing data

According to Lai and Leu (2017), data pre-processing uses a desired analytical method to obtain

good quality and efficient data analysis results by eliminating inconsistent, abnormal, and erroneous data. Several activities are performed during the pre-processing step, including checking missing data, inequalities, tokenisation, stop words, and stemming.

#### 2.3.1. Checking missing data

After the data is imported, the next step is to ensure that no data (values) are missing from each column/variable because they can negatively affect the classification prediction results (Haq et al., 2019). In the research presented in this paper, the function used to check the missing data is df.isnull().sum().

#### 2.3.2. Checking inequalities

This step was performed to prepare the training set data and test set data for classification steps. The purpose is to check if there is an inequality in the number of labels on the "product design criteria" and "sentiment of the review" variables. This step is important because the classification accuracy is strongly influenced by the balance of the data amount. The model trained using unequal data has low accuracy because it tends to predict data to classes with more numbers. A large amount of data cannot improve the classification results if the problem of data inequality is not addressed (Li et al., 2011; Shen et al., 2019). The function used to handle inequality is df['Criteria'].value_counts() and df['Sentiment'].value_counts().

#### 2.3.3. Tokenisation, stop-word removal, and stemming

Tokenisation is dividing text into a series of words called tokens. Stemming is changing words into their basic form by removing affixes. Eliminating affixes can increase recall (the model's ability to identify true positives) (Issac & Jap, 2009). Stop-word removal is removing words without a significant meaning in the text (Khedr et al., 2017). In this proposed framework, pre-processing data is performed using Python. The stop-word removal activity used the stop-word library from the "nltk" package in Python. NLTK (Natural Language Toolkit) is a platform for processing human language data in Python (Bird et al., 2009). The stemming process is done by defining a function using Porter Stemmer from NLTK and then applying that function to the review text. Porter Stemmer is a stemming algorithm that removes common morphological endings and inflections from words in English (Issac & Jap, 2009).

### 2.4. Feature extraction

Feature extraction takes a list of words from text data and converts it into a set of features that can be used by the classifier (Waykole & Thakare, 2018). The feature extraction technique used was frequency-inverse document frequency (TF-IDF). TF-IDF measures the relevance of a term in a document. Where $w_{i,j}$ is the weight of term i in document j, N is the number of documents, $tf_{i,j}$ is the frequency of term i in document j, and $df_i$ is the frequency of documents of term i, then the weighting of terms in TF-IDF can be formulated in the following equation (Zhang et al., 2011):

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right). \quad (1)$$

The first step is to use the "CountVectorizer()" function from the "sklearn" library. This function converts a text into a vector based on the frequency of each word that appears in the text. After converting into vector form, the next step is to perform TF-IDF using the "TfidfTransformer()" function from the "sklearn" library. The main purpose for using TF-IDF in calculating the number of tokens that appear in a document compared to the "raw frequency" is to reduce the impact of tokens that appear frequently in documents. These frequent tokens are empirically less informative when compared to tokens that appear in small fractions (Pedregosa et al., 2011). After the data is transformed into a TF-IDF array, the next step is to divide the data into two data sets: training data and test data. The training data is used to train the created classification model. The distribution of the dataset uses a ratio of 80 % for training data and 20 % for test data based on suggestions by Rácz et al. (2021).

### 2.5. Classification

After the feature extraction, the classification technique is applied based on "product design criteria" and "sentiment of the review". Because the technique proposed in the research presented in this paper is supervised learning, the data is divided into a training set and a test set. Usually, there is a problem related to data inequalities leading to data discrepancy when deciding which data are categorised as a training set and a test set. The data discrepancy makes the classification model biased towards a larger number of classes, resulting in poor model performance (Guo et al., 2019). The Synthetic Minority Oversampling Technique (SMOTE) method is used to overcome this inequality. Elreedy

and Atiya (2019) described the SMOTE stages as follows:

- For each $X_0$ pattern of the minority class, do the following:
  - Choose one of the K closest neighbours to X (nearest neighbours), who are also a minority class.
  - Create a new pattern Z at a random point on the line segment connecting the pattern and the selected neighbours. Z is formulated as:

$$Z = X_0 + \omega(X - X_0), \quad (2)$$

    where $\omega$ is a random variable that is uniformly distributed U [0,1].
    The SMOTE method is performed using the "SMOTE()" function in the "imblearn" library
- After the data discrepancy is resolved, sentiment classification and criteria are carried out on the training data set.

The proposed framework compared three supervised Machine Learning classification techniques: Naïve Bayes classification (Kang & Park, 2014; Povoda, 2016), Logistic Regression (Pranckevičius & Marcinkevičius, 2017; Shah et al., 2020; Wang et al., 2017), and Support Vector Machine (SVM) (Hadi et al., 2018; Jiang et al., 2013; Singh et al., 2019; Tan et al., 2009). After the classification model is trained, the results are validated using a test data set to run the trained model. The validity of the model is measured by the level of accuracy given by the model using the test data set. The model of the classification technique with the best performance was chosen as the reference classifier model.

### 2.6. Interpreting the result from the categorisation requirement process

In this step, the result in the categorisation requirement process started from getting the customer review, data labelling, feature extraction and classification, interpreted to give an insight that can be used by the company to identify the criteria considered by the consumer in buying a certain type of product.

The result is presented using a data visualisation technique so that it is easier for the company to understand.

## 3. Case study: customer reviews of women's clothing in e-commerce

This section discusses the application of the framework proposed in a case study. The selected case study is a customer review of women's clothing on an e-commerce site.

### 3.1. Getting a customer review by web scraping techniques

To illustrate the applicability of the proposed framework, data on women's clothing reviews from an e-commerce site are used. The data is downloaded from www.kaggle.com based on data by Agarap (2018). This data consists of ten variables. Among the ten variables provided in the data set, the research presented in this paper uses the "Review Text" variable, which contains reviews from each user as shown in Table 2.

### 3.2. Data labelling

There are two labels used to mark the review text by the user: "product design criteria" and "sentiment of the review". Each label has four and three classes, respectively. The label "sentiment of the review" has three classes: "positive", "neutral", and "negative", indicating whether a review is positive, neutral, or negative. Four classes of the "product design criteria" are taken from Eckman et al. (1990) on the four main criteria that consumers consider when deciding to buy women's clothing. The four criteria can be seen in Table 3.

The class of the "product design criteria" label is determined by looking at the tendency of user reviews towards one of the sub-criteria listed in Table 3. If the review meets one of the sub-criteria, then the class assigned to the label "product design criteria" in the review is the criteria related to the sub-criteria.

For example, the review "This shirt is very flattering to all due to the adjustable front tie. It is the perfect length to wear with leggings and it is sleeveless, so it pairs well with any cardigan. love this shirt!!!" discusses how the shirts purchased by users are very suitable when paired with leggings and cardigans. This relates to the "matching" sub-criteria, so the class

Tab. 2. Description column name

| No | Column name | Description |
|---|---|---|
| 1 | Clothing ID | An integer variable that refers to one of the items under review |
| 2 | Age | Age of the user who gave the review |
| 3 | Title | The title of the user-generated review |
| 4 | Review Text | Contents of user-generated reviews |
| 5 | Rating | The value assigned by the user for the product being reviewed. The range of values given is from 1 (worst) to 5 (best) |
| 6 | Recommended IND | A variable that states whether the user recommends the product or not |
| 7 | Positive Feedback Count | A variable that shows the number of other users who consider the reviews made positive |
| 8 | Division Name | Name of the division of the product under review |
| 9 | Department Name | Name of the department of the product under review |
| 10 | Class Name | The class name of the product under review |
| 11 | Clothing ID | An integer variable that refers to one of the items under review |

Tab. 3. Criteria for women's clothing

| Criteria | Sub-Criteria | Definition |
|---|---|---|
| Aesthetic | Colour/pattern | Relates to the colour, print, or visual pattern of clothing |
| | Styling | Associated with the design of clothing that includes the fashionable, style, or individual preference for a type of clothing |
| | Fabric | Relates to the material and content of the fabric used to make clothes |
| | Uniqueness | Associated with the uniqueness, unusualness, and rarity of clothing |
| | Appearance | Relates to how the clothes look to the user: attractive vs. not attractive appearance |
| Usefulness | Versatility | Relates to the adaptability of clothing to various uses or situations |
| | Matching | Relates to the suitability of clothes when paired with other clothes |
| | Appropriateness | Relates to the suitability of clothing for a particular social and occupational environment |
| | Utility | Related to the ability of clothing to meet the needs of certain clothes from its users |
| Performance and quality | Fit | Relates to the suitability of clothes to the body shape |
| | Comfort | Relates to how the clothes and the material of the clothes are perceived by the wearer |
| | Care | Relates to how the clothes are taken care of by the wearer |
| | Workmanship | Relates to the superiority of construction/manufacture or material of clothing |
| Extrinsic | Price | Relates to the price of the clothes |
| | Brand | Associated with the name of the clothing maker or brand of the clothing |
| | Competition | Relates to the availability of the same type of clothing in other stores |

for the "product design criteria" label assigned to the review is "usefulness". The class assigned to the "sentiment of the review" label was "positive" because the review was positive, marked by the words "This shirt is very flattering…" and "… love this shirt!!!".

### 3.3. Pre-processing data

Data pre-processing activity was performed before proceeding to the feature extraction steps. This step was conducted to process the data to the appropriate form for further analysis. Pre-processing in the research proposed in this paper is mostly done using Python through the Google Collaboratory/Google Colab platform. This paper does not explain tokenisation, stop words, and stemming in detail because those are common steps in text mining methods.

#### 3.3.1. Checking missing data

The function used to check the missing data is df.isnull().sum(). The result is presented in Fig. 2.

```
Review Text    0
Kriteria       0
Sentiment      0
dtype: int64
```

Fig. 2. Result from checking missing values

### 3.3.2. Handling inequality

The functions used to handle inequality is df['Criteria'].value_counts() and df['Sentiment'].value_counts(). The result is presented in Figs. 3 and 4.

```
Performance and quality    1310
Aesthetic                   722
Usefulness                  135
Extrinsic                    42
```

Fig. 3. Check the balance data for the "product design criteria" variable

```
Positive    1611
Neutral      300
Negative     298
```

Fig. 4. Check the balance of data for the "sentiment of the review" variable

Based on Figs. 3 and 4, there is a disparity in the number of classes on both the "product design criteria" label and the "sentiment of the review" label. The "product design criteria" label is dominated by the "performance and quality" class, followed by the "aesthetic" class. "Sentiment of the review" is dominated by the "positive" class.

### 3.3.3. Tokenisation, stop-word removal, and stemming

Several activities performed in this step include the removal of punctuation, stop words and stemming.

### 3.4. Feature extraction

As mentioned in Section 3.3.2, there is an imbalance in the number of classes from the labels "product design criteria" and "sentiment of the review". Therefore, the SMOTE method was used to address this inequality. By overcoming data inequality, it is hoped that there will be no overfitting in the classification model that will be made.

The amount of data for each class before the data inequality is overcome can be seen in Fig. 5.

Based on Fig. 5, the amount of data in each class of the two labels becomes the same after applying the SMOTE method.

### 3.5. Classification

In this step, the proposed multilabel classification technique used "product design criteria" and "sentiment of the review", which has not been done in three supervised machine learning techniques for classification. Multilabel classification technique is based on "product design criteria" and "sentiment of the review". This sub-section discusses the performance of the selected Machine Learning (ML) classification techniques: Support Vector Machine, Naïve Bayes, and Logistic Regression.

### 3.5.1. Support Vector Machine (SVM)

The SVM model is imported from the "sklearn" library. Data training is done using the "model.fit()" function from the "sklearn" library. The variable "x_train", which contains the review text from consumers, is used as a predictor variable, and the variable "y_train" is used as the response variable.

The response variables are the labels "product design criteria" and "sentiment of the review" so that the model training is carried out once each for both labels. After the model is trained, the next step is to

```
Before OverSampling, counts of label Performance and quality is: 1041
Before OverSampling, counts of label Aesthetic is: 571
Before OverSampling, counts of label Usefulness is: 115
Before OverSampling, counts of label Extrinsic is: 40
Total number of rows: 1767

Before OverSampling, counts of label is Positive: 1286
Before OverSampling, counts of label is Negative: 237
Before OverSampling, counts of label is Neutral: 244
Total number of rows: 1767
```

Fig. 5. Total data before inequality is overcome

calculate the accuracy of the model using the test data set. The number of test data from the split results is 442 for the "product design criteria" label and 302 for "sentiment of the review". Accuracy is calculated by comparing the result label predicted by the model with the actual label. Prediction of the SVM model for the "product design criteria" label can be seen in Fig. 6.

From Fig. 6, the results of the accuracy measurement of the SVM model for the "product design criteria" label give an accuracy value of 62.2 %. The f1-score indicates what percentage of the predicted label corresponds to the actual label. Furthermore, the prediction of the SVM model for the "sentiment of the review" label can be seen in Fig. 7.

From Fig. 7, the accuracy obtained from the model is 76.0 % with precision and recall values for the "negative", "neutral", and "positive" classes are 0.56 and 0.57; 0.26 and 0.25; and 0.88 and 0.88, respectively. An example of a neutral tone of review can be seen in Table 4.

Of the 61 "negative" data, 35 were correctly predicted, while 15 and 11 were predicted as "neutral" and "positive", respectively. It can be concluded that classes with a high support value (amount of class data) tend to have a higher f1-score compared to other classes.

Accuracy 0.6221719457013575

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Aesthetic | 0.53 | 0.54 | 0.54 | 151 |
| Extrinsic | 0.50 | 0.50 | 0.50 | 2 |
| Performance and quality | 0.72 | 0.71 | 0.71 | 269 |
| Usefulness | 0.10 | 0.10 | 0.10 | 20 |
| accuracy | | | 0.62 | 442 |
| macro avg | 0.46 | 0.46 | 0.46 | 442 |
| weighted avg | 0.63 | 0.62 | 0.62 | 442 |

Fig. 6. Label classification "product design criteria" using the SVM model

Accuracy 0.7601809954751131

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Negative | 0.56 | 0.57 | 0.56 | 61 |
| Neutral | 0.26 | 0.25 | 0.25 | 56 |
| Positive | 0.88 | 0.88 | 0.88 | 325 |
| accuracy | | | 0.76 | 442 |
| macro avg | 0.57 | 0.57 | 0.57 | 442 |
| weighted avg | 0.76 | 0.76 | 0.76 | 442 |

Fig. 7. Label classification "sentiment of the review", the SVM model

Tab. 4. Example of review "neutral"

| Review | Rating | "Product design criteria" | "Sentiment of the review" |
|---|---|---|---|
| "this is a cute top that can transition easily from summer to fall. it fits well, nice print and it's comfortable. i tried this on in the store but did not purchase it because the color washed me out. this is not the best color for a blonde. would look much better on a brunette. if this was a different color i most likely would have purchased it." | 4 | Aesthetic | Neutral |

### 3.5.2. Naïve Bayes

The data training steps with the Naïve Bayes model are the same as the steps performed in the SVM model. The Naïve Bayes model is imported from the "sklearn" library. Data training is done using the "model.fit()" function from the "sklearn" library. The prediction of the "product design criteria" label using the Complement Naïve Bayes model can be seen in Fig. 8.

The accuracy obtained from the Naïve Bayes model is 50.6 %. The confusion matrix of the Naïve Bayes model for the prediction of the "product design criteria" label shows that from 269 "performance and quality" data, 152 are correctly predicted. Furthermore, the prediction of the Complement Naïve Bayes model for the label "sentiment of the review" can be seen in Fig. 9.

From the results of the classification of the label "sentiment of the review" by the Complement Naïve Bayes model, an accuracy value of 72.3 % was obtained.

Just like the previous SVM model, the low f1-score in the "neutral" class can be caused by neutral reviews that tend to discuss positive and negative aspects of clothing.

### 3.5.3. Logistic Regression

The Logistic Regression model is imported from the "sklearn" library. Data training is done using the "model.fit()" function from the "sklearn" library. According to the "sklearn" documentation, the model uses Multinomial Logistic Regression if more than two classes are detected (Pedregosa et al., 2011).

Accuracy 0.5067873303167421

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Aesthetic | 0.50 | 0.40 | 0.44 | 151 |
| Extrinsic | 0.03 | 0.50 | 0.06 | 2 |
| Performance and quality | 0.73 | 0.57 | 0.64 | 269 |
| Usefulness | 0.14 | 0.55 | 0.22 | 20 |
| accuracy | | | 0.51 | 442 |
| macro avg | 0.35 | 0.50 | 0.34 | 442 |
| weighted avg | 0.62 | 0.51 | 0.55 | 442 |

Fig. 8. Result classification label "product design criteria", Naïve Bayes model

Accuracy 0.7239819004524887

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Negative | 0.50 | 0.54 | 0.52 | 61 |
| Neutral | 0.23 | 0.34 | 0.28 | 56 |
| Positive | 0.91 | 0.82 | 0.87 | 325 |
| accuracy | | | 0.72 | 442 |
| macro avg | 0.55 | 0.57 | 0.55 | 442 |
| weighted avg | 0.77 | 0.72 | 0.74 | 442 |

Fig. 9. Label classification "sentiment of the review", Naïve Bayes model complement

Accuracy 0.6402714932126696

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Aesthetic | 0.56 | 0.58 | 0.57 | 151 |
| Extrinsic | 0.17 | 0.50 | 0.25 | 2 |
| Performance and quality | 0.75 | 0.70 | 0.72 | 269 |
| Usefulness | 0.22 | 0.30 | 0.26 | 20 |
| accuracy | | | 0.64 | 442 |
| macro avg | 0.42 | 0.52 | 0.45 | 442 |
| weighted avg | 0.66 | 0.64 | 0.65 | 442 |

Fig. 10. Result of label classification "product design criteria", Logistic Regression model

Accuracy 0.7805429864253394

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Negative | 0.62 | 0.64 | 0.63 | 61 |
| Neutral | 0.32 | 0.36 | 0.34 | 56 |
| Positive | 0.91 | 0.88 | 0.89 | 325 |
| accuracy | | | 0.78 | 442 |
| macro avg | 0.61 | 0.63 | 0.62 | 442 |
| weighted avg | 0.79 | 0.78 | 0.79 | 442 |

Fig. 11. Model result of classification label "sentiment of the review", Logistic Regression model

The prediction of the "product design criteria" label from the Logistic Regression model can be seen in Fig. 10.

The accuracy value obtained from the Logistic Regression model is 64 %.

Predictions of the Logistic Regression model for the "sentiment of the review" label can be seen in Fig. 11.

The accuracy value obtained from the Logistic Regression model for the "sentiment of the review" label is 78 %.

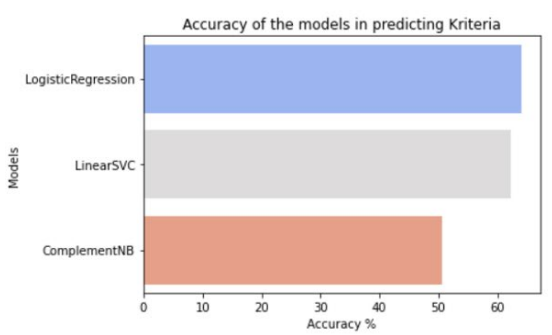
### 3.5.4. Comparison of classification results

After the performance of the three models is measured, the next step is to choose the best model that is going to be used to predict the "product design criteria" and "sentiment of the review" labels from the next customer review. The performance of the three models can be seen in Fig. 12.

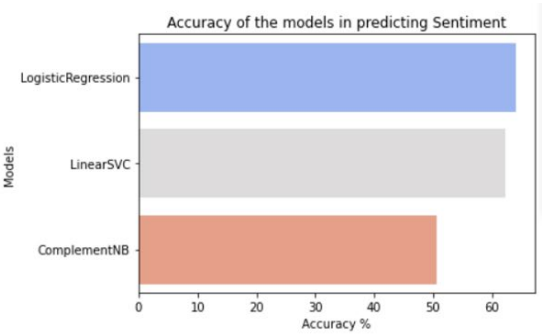
According to Hair et al. (2014), a classification model has good accuracy if the classification accuracy value is 25 % higher than the theoretical probability. The theoretical probability of a review being classified into one of the classes on the "product design criteria" label is 25 % (one of four class choices) and 33.3 % for the "sentiment of the review" label (one of three class choices). Thus, the minimum accuracy value that must be achieved by the model for the "product design criteria" label is 31 %, and the "sentiment of the review" label is 42 %. All three models provide accurate results exceeding the minimum limit.

Based on Fig. 12 above, the model with the highest accuracy for both labels is the Logistic Regression model, with an accuracy value for the "product design criteria" label of 64 % and for the "sentiment of the review" label of 78 %. So, in this case, the Logistic Regression model was chosen as the model used to predict the labels "product design criteria" and "sentiment of the review".

As demonstrated by the three models above, the most common class grouping errors occurred in the "performance and quality" and "aesthetic" classes. This implies that the terminology used in the review of the two classes intersects. From the description of




| Models | Accuracy |
|---|---|
| LogisticRegression | 64.027149 |
| LinearSVC | 62.217195 |
| ComplementNB | 50.678733 |




| Models | Accuracy |
|---|---|
| LogisticRegression | 78.054299 |
| LinearSVC | 76.018100 |
| ComplementNB | 72.398190 |

Fig. 12. Performance of the three supervised classification technique models

Tab. 5. Example of a review of "aesthetic" and "performance and quality" criteria

| Review text | Product design criteria |
|---|---|
| Cute little dress fits tts. it is a little high waisted. good length for my 5'9 height. i like the dress, i'm just not in love with it. i dont think it looks or feels cheap. it appears juts as pictured | Aesthetic |
| Loved the material, but i didn't really look at how long the dress was before i purchased both a large and a medium. im 5'5" and there was atleast 5" of material at my feet. the gaps in the front are much wider than they look. felt like the dress just fell flat. both were returned. im usually a large and the med fit better. 36d 30 in jeans | Performance and quality |

each sub-criteria for the "performance and quality" and "aesthetic" classes, there is a possibility of similarities in the terminology used when the review is classified into the "styling" sub-criteria from the "aesthetic" criteria and the "fit" sub-criteria from the "aesthetic" criteria. "performance and quality". An example of a review of "aesthetic" and "performance and quality" criteria can be seen in Table 5.

From Table 5, the review text of the two criteria both mentions "size", but in context one review discusses appearance and the other review discusses the fit of the clothes being reviewed.

## 4. Result and discussion

The proposed framework presented in this paper can result in the requirement categorisation of customer needs based on "product design criteria" and "sentiment of the review" labels. The result of the requirement categorisation using the two labels is presented in Fig. 13.

The results of this classification can be used as a reference in determining the direction of product development. One such direction is to create products that can resolve or answer customer complaints about a criterion by looking at customer reviews with negative sentiments. Determining the direction of product development is hoped to reduce product development costs because it can save time for testing hypotheses and reduce the cost of design changes (cost of change) in later phases of product development (Folkestad & Johnson, 2001; Pedersen et al., 2016).

Based on Fig. 13, "performance and quality" are the criteria with the most negative sentiments. Decision makers can see the most used negative words in reviews about "performance and quality". Of the twenty words, the type of specific clothing that is most often mentioned is the type of a dress or a dress. When explored further, it is found that most of the words related to a dress are related to its size or fit.

Based on customer reviews about a dress presented in Table 6, negative responses regarding the "performance and quality" criteria of the dress are the

| | | | |
|---|---|---|---|
| **0** | I love the color of this dress. it is not flat... | Aesthetic | Negative |
| **1** | I received the vest and it was pretty much as ... | Aesthetic | Neutral |
| **2** | This dress is really beautiful, cheerful, the ... | Aesthetic | Neutral |
| **3** | I'll admit that i often believe you pay for th... | Performance and quality | Positive |
| **4** | I recently purchased this dress in the yellow,... | Aesthetic | Positive |
| **...** | ... | ... | ... |
| **12469** | I was surprised at the positive reviews for th... | Aesthetic | Negative |
| **12470** | So i wasn't sure about ordering this skirt bec... | Aesthetic | Positive |
| **12471** | I was very happy to snag this dress at such a ... | Extrinsic | Negative |
| **12472** | This fit well, but the top was very see throug... | Performance and quality | Positive |
| **12473** | I bought this dress for a wedding i have this ... | Performance and quality | Neutral |

Fig. 13. Result of requirement categorisation steps based on "product design criteria" and "sentiment of the review" labels

Tab. 6. Example of review of "aesthetic" criteria; "performance and quality"

| |
|---|
| Just piping in here -- ordered my usual size of small petite. the slip that came with the dress is about a size negative zero, it could hardly squeeze over my body and the dress itself is a bright pale aqua and it is a shift. and because of the smocking it hangs very strangely. i think it looks very cheap and is ill fitting. i would say if you are a person on which shift dresses look awesome you might like this, but mind the size of the slip in the petites range, and also it is aqua. |
| I'm a fan of yumi kim dresses and consistently wear size small. so, i ordered this in a small, and the top was a little loose, but it was way too short. there's no way i would feel comfortable wearing this in public. i'm 5'7", and the hem was higher than mid-thigh on me. the waist is also billowy and has no defined shape. |
| This is a pretty dress. i bought a size 0 and the body of the dress fit, however, the sleeves were too long. had to return, as the loose and longer sleeves made the dress look too big on me. |
| I was really excited to receive this dress. it's a fun concept and different from anything else i have. however, the dress was way too short on me. i'm 5 ft 6 in., and the size medium was way too short. given the thin nature of the material, i did not feel comfortable keeping the dress, even for wearing it outside the office. i will have to return this dress. |
| Ordered this dress in an xs, that is the size i usually wear. this runs small, but not only that it is the way it is cut, very small through the hips, thighs and legs. hard to walk in it. the material felt kind of cheap to me. i thought it looked really good on the model, but got it home and didn't like it at all! it went back! |
| This dress looks cute online, but it is enormous. i bought a small, but it looked more like a 3x plus size. i tried it on, just in case, and honestly it looked like a great dress for a clown. this dress must have had two feet of fabric pinned behind the model in the photo. |

discrepancy between the clothing sizes listed and the actual clothing sizes.

Thus, decision-makers can determine the direction of product development of dresses with sizes that suit most customers' body sizes. In addition, the framework proposed in the research presented in this paper can improve the process of product development, especially during the conceptual stage.

Based on the results of the case study in Section 3, the proposed framework can fill the gaps presented by Shabestari et al. (2019). The first contribution is that the proposed framework can show the success of using supervised learning techniques in the initial stages of product development, where the use of supervised learning techniques for classification is said to be better than unsupervised learning techniques (Abad et al., 2017; Edwards et al., 2021). The second contribution is that in the case study, a comparison was made of the three classification techniques, in which, for this case, it can be concluded that the Logistic Regression technique shows better accuracy than the other two techniques.

Meanwhile, the results demonstrate that the proposed framework can produce sentiment output, which is easier to use as a basis in the product development process. This is different from other techniques, as stated by Ireland and Liu (2018), where although it is easy to obtain customer preference data using sentiment analysis techniques, it is still difficult to use the results as insights in the product development process. This proposed framework is different because of the two label kinds, "product design criteria" and "sentiment of the review", and the resulting output, which is similar to Fig. 13, of the product features can be broken down, and the VoC for these features can be raised. This way, the output of sentiment analysis can be directly applied as a basis in the product development process.

To apply the proposed framework for other types of products, it is necessary to determine product criteria and sub-criteria (product design criteria), which is similar to Table 3 in the case study above, referring to the processes and standards in the product to be developed. With adequate product design criteria, the proposed framework can obtain VoC, which facilitates the product development process.

## Conclusions

Product development can generally be grouped into three major stages: conceptual design, prototyping, and product launch. Voice of Customer (VoC) is especially needed during the product development stages in the conceptual design. One way to get VoC is through customer reviews. Based on the set of data on women's clothing reviews from an e-commerce site downloaded from www.kaggle.com based on data by Agarap (2018), it can be proved (Section 3) that the proposed framework (Fig. 1) can be applied. The steps of the product development process, especially in the requirement categorisation stage, can be performed using the framework presented in this research, i.e., data labelling, data pre-processing, feature extraction, and classification. The research presented in this paper used a supervised multilabel classification technique of Logistic Regression. The framework proposed in this study can also be used in general for all types of products if the product criteria

are well-defined and can be used to label customer reviews. In addition, the research presented in this paper also answers one of the concerns in the categorisation of needs raised by Shabestari et al. (2019), namely, the unclear rules or main attributes of a requirement that make these needs fall into certain categories.

## Literature

Abad, Z. S. H., Karras, O., Ghazi, P., Glinz, M., Ruhe, G., & Schneider, K. (2017). What Works Better? A Study of Classifying Requirements. *Proceedings of 2017 IEEE 25th International Requirements Engineering Conference, RE 2017*, 496-501. doi: 10.1109/RE.2017.36

Agarap, A. F. (2018). Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network (RNN). *arXiv preprint arXiv*:1805.03687. Retrieved from http://arxiv.org/abs/1805.03687

Aguwa, C., Olya, M. H., & Monplaisir, L. (2017). Modeling of fuzzy-based voice of customer for business decision analytics. *Knowledge-Based Systems*, *125*, 136-145. doi: 10.1016/j.knosys.2017.03.019

Bhuiyan, N. (2011). A framework for successful new product development. *Journal of Industrial Engineering and Management*, *4*(4), 746-770. doi: 10.3926/jiem.334

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Booz, A., & Hamilton (1982). *New Products Management for the 1980s*. Booz, Allen and Hamilton, Tysons Corner.

Broucke, S.V., & Baesens, B. (2018). *Practical Web Scraping for Data Science: Best Practices and Examples with Python*. Apress.

Büyüközkan, G., & Arsenyan, J. (2012). Collaborative product development: A literature overview. *Production Planning and Control*, *23*(1), 47-66. doi: 10.1080/09537287.2010.543169

Carter, M. P. (2015). *Creation and validation of a best practice new product development process assessment tool for industrial practitioners*. Doctoral dissertation, Indiana State University.

Cooper, R. G. (2001). *Winning at New Products: Accelerating the Process from Idea to Launch* (3rd ed.). Perseus Books.

Devi, D. V. N., Kumar, C. K., & Prasad, S. (2016). A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine. *Proceedings of 2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 3-8. doi: 10.1109/IACC.2016.11

Di Benedetto, C. A. (1999). Identifying the key success factors in new product launch. *Journal of Product Innovation Management*, *16*(6), 530-544. doi: 10.1111/1540-5885.1660530

Eckman, M., Damhorst, M. L., & Kadolph, S. J. (1990). Toward a Model of the In-Store Purchase Decision Process: Consumer Use of Criteria for Evaluating Women's Apparel. *Clothing and Textiles Research Journal*, *8*(2), 13-22. doi: 10.1177/0887302X9000800202

Edwards, A. S., Kaplan, B., & Jie, T. (2021). A Primer on Machine Learning. *Transplantation*, *105*(4), 699-703. doi: 10.1097/TP.0000000000003316

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32-64. doi: 10.1016/j.ins.2019.07.070

Folkestad, J. E., & Johnson, R. L. (2001). Resolving the conflict between design and manufacturing: Integrated Rapid Prototyping and Rapid Tooling (IRPRT). *Journal of Industrial Technology*, *17*(4), 1-7.

Gonçalves, P., Benevenuto, F., Araujo, M., & Cha, M. (2013). Comparing and Combining Sentiment Analysis Methods Categories and Subject Descriptors. *Proceedings* of the *first ACM conference on Online social networks (COSN '13),* 27-38. doi: 10.1145/2512938.2512951

Guo, S., Liu, Y., Chen, R., Sun, X., & Wang, X. (2019). Improved SMOTE Algorithm to Deal with Imbalanced Activity Classes in Smart Homes. *Neural Processing Letters*, *50*(2), 1503-1526. doi: 10.1007/s11063-018-9940-3

Gupta, M., & Sebastian, S. (2018). Framework to analyze customer's feedback in smartphone industry using opinion mining. *International Journal of Electrical and Computer Engineering*, *8*(5), 3317-3324. doi: 10.11591/ijece.v8i5.pp3317-3324

Hadi, W., Al-Radaideh, Q. A., & Alhawari, S. (2018). Integrating associative rule-based classification with Naïve Bayes for text classification. *Applied Soft Computing Journal*, *69*, 344-356. doi: 10.1016/j.asoc.2018.04.056

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate Data Analysis* (7th ed.). Pearson Education Limited.

Haq, A. U., Li, J., Khan, J., Memon, M. H., Parveen, S., Raji, M. F., Akbar, W., Ahmad, T., Ullah, S., Shoista, L., & Monday, H. N. (2019). Identifying The Predictive Capability of Machine Learning Classifiers For Designing Heart Disease Detection System. *Proceedings of the 16th International Computer Conference on Wavelet Active Media Technology and Information Processing,* 130-138. doi: 10.1109/ICCWAMTIP47768.2019.9067519.

Huang, S., Peng, W., Li, J., & Lee, D. (2013). Sentiment and Topic Analysis on Social Media: A Multi-Task Multi-Label Classification Approach. *Proceedings of the 5th Annual ACM Web Science Conference*, 172-181. doi: 10.1145/2464464.2464512

Ireland, R., & Liu, A. (2018). Application of data analytics for product design: Sentiment analysis of online product reviews. *CIRP Journal of Manufacturing Science and Technology*, *23*, 128-144. doi: 10.1016/j.cirpj.2018.06.003

Issac, B., & Jap, W. J. (2009). Implementing spam detection using Bayesian and porter stemmer keyword stripping approaches. *Proceedings of TENCON 2009 IEEE Region 10 Conference,* 1-5. doi: 10.1109/TENCON.2009.5396056

Jeong, B., & Yoon, J. (2016). Identifying product opportunities using topic modeling and sentiment analysis of social media data. *Proceedings of the 17th Asia Pacific Industrial Engineering and Management System Conference,* Paper 208.

Jiang, L., Cai, Z., Zhang, H., & Wang, D. (2013). Naive Bayes text classifiers: A locally weighted learning approach. *Journal of Experimental & Theoretical Artificial Intelligence*, *25*(2), 273-286. doi: 10.1080/0952813X.2012.721010

Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications*, *41*(4-1), 1041-1050. doi: 10.1016/j.eswa.2013.07.101

Kapucugil Ikiz, A., & Özdağoğlu, G. (2015). Text Mining as a Supporting Process for VoC Clarification. *Alphanumeric Journal*, *3*(1), 25-40.

Kaulio, M. A. (1998). Customer, consumer, and user involvement in product development: A framework and a review of selected methods. *Total Quality Management*, *9*(1), 141-149. doi: 10.1080/0954412989333

Khedr, A. E., Salama, S. E., & Yaseen, N. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, *9*(7), 22-30. doi: 10.5815/ijisa.2017.07.03

Kim, Y. H., Park, S. W., & Sawng, Y. W. (2016). Improving new product development (NPD) process by analyzing failure cases. *Asia Pacific Journal of Innovation and Entrepreneurship*, *10*(1), 134-150. doi: 10.1108/APJIE-12-2016-002

Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*. Retrieved from http://arxiv.org/abs/1507.00955

Kornish, L. J., & Hutchison-Krupat, J. (2017). Research on Idea Generation and Selection: Implications for Management of Technology. *Production and Operations Management*, *26*(4), 633-651. doi: 10.1111/poms.12664

Kreuzbauer, R., & Malter, A. J. (2005). Embodied cognition and new product design: Changing product form to influence brand categorization. *Journal of Product Innovation Management*, *22*(2), 165-176. doi: 10.1111/j.0737-6782.2005.00112.x

Lai, S. T., & Leu, F. Y. (2017). Data preprocessing quality management procedure for improving big data applications efficiency and practicality. *Lecture Notes on Data Engineering and Communications Technologies*, *2*, 731-738. doi: 10.1007/978-3-319-49106-6_73

Lee, T. Y., & Bradlow, E. T. (2011). Automated Marketing Research Using Online Customer Reviews. *Journal of Marketing Research*, *48(5)*, 881-894. doi: 10.1509/jmkr.48.5.881

Li, S., Wang, Z., Zhou, G., & Lee, S. Y. M. (2011). Semi-Supervised Learning for Imbalanced Sentiment Classification. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence,* 1826-1831.

Malviya, S., Tiwari, A. K., Srivastava, R., & Tiwari, V. (2020). Machine learning techniques for sentiment analysis: A review. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *12*(02), 72-78. doi: 10.18090/samriddhi.v12i02.03

Micu, A., Micu, A. E., Geru, M., & Lixandroiu, R. C. (2017). Analyzing user sentiment in social media: Implications for online marketing strategy. *Psychology & Marketing*, *34*(12), 1094-1100. doi: 10.1002/mar.21049

Misopoulos, F., Mitic, M., Kapoulas, A., & Karapiperis, C. (2014). Uncovering customer service experiences with Twitter: The case of airline industry. *Management Decision*, *52*(4), 705-723. doi: 10.1108/MD-03-2012-0235

Mulay, R., & Khanna, V. T. (2017). A Study on the Relationship between the Voice of Customer with the Cost of Quality in Processes of Professional Higher Education Institutions. *South Asian Journal of Management*, *24*(4), 55.

Nazari-Shirkouhi, S., & Keramati, A. (2017). Modeling customer satisfaction with new product design using a flexible fuzzy regression-data envelopment analysis algorithm. *Applied Mathematical Modelling*, *50*, 755-771. doi: 10.1016/j.apm.2017.01.020

Park, J., Lee, H., Lee, J. H., & Suh, H. W. (2018). Feature-based sentiment word selection and rating for system design. *Journal of Industrial Electronics Technology and Application*, *1*(4), 54-57.

Park, Y. E., & Alenezi, M. (2018). Predicting the popularity of Saudi multinational enterprises using a data mining technique. *Journal of Management Information and Decision Science*, *21*(1), 1-15.

Pedersen, S. N., Christensen, M. E., & Howard, T. J. (2016). Robust design requirements specification: a quantitative method for requirements development using quality loss functions. *Journal of Engineering Design*, *27*(8), 544-567. doi: 10.1080/09544828.2016.1183163

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Varoquaux, G., Gramfort, A., Thirion, B., Dubourg, V., Passos, A., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Pienaar, C., van der Lingen, E., & Preis, E. (2019). A framework for successful new product development. *South African Journal of Industrial Engineering*, *30*(3), 199-209. doi: 10.7166/30-3-2239

Pinquié, R., Véron, P., Segonds, F., & Croué, N. (2018). A requirement mining framework to support complex sub-systems suppliers. *Procedia CIRP*, *70*, 410-415. doi: 10.1016/j.procir.2018.03.228

Povoda, L., Burget, R., & Dutta, M. K. (2016). Sentiment analysis based on support vector machine and big data. *Proceedings of the 39th International Conference on Telecommunications and Signal Processing*, 543-545. doi: 10.1109/TSP.2016.7760939

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*, *5*(2), 221-232. doi: 10.22364/bjmc.2017.5.2.05

Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*, *26*(4), 1111. doi: 10.3390/molecules26041111

Samuel, J., Ali, G. G. M. N., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). COVID-19 public sentiment insights and machine learning for tweets classification. *Information*, *11*(6), 314. doi: 10.3390/info11060314

Schulze, A., & Hoegl, M. (2006). Knowledge creation in new product development projects. *Jour-*

*nal of Management*, *32*(2), 210-236. doi: 10.1177/0149206305280102

Shabestari, S. S., Herzog, M., & Bender, B. (2019). A survey on the applications of machine learning in the early phases of product development. *Proceedings of the Design Society: International Conference on Engineering Design*, 2437-2446. doi: 10.1017/dsi.2019.250

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, *5*(1), 12. doi: 10.1007/s41133-020-00032-0

Shen, J., Baysal, O., & Shafiq, M. O. (2019). Evaluating the Performance of Machine Learning Sentiment Analysis Algorithms in Software Engineering. *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 1023-1030. doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00185

Shukri, S. E., Yaghi, R. I., Aljarah, I., & Alsawalqah, H. (2015). Twitter sentiment analysis: A case study in the automotive industry. *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT 2015*, 1-5. doi: 10.1109/AEECT.2015.7360594

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. *International Conference on Automation, Computational and Technology Management (ICACTM)*, 593-596. doi: ICACTM.2019.8776800

Šperková, L. (2019). Qualitative Research on Use of Voice of Customer in Czech Organisations. *Journal of Systems Integration*, *10*(2), 9-19.

Tan, S., Cheng, X., Wang, Y., & Xu, H. (2009). Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. *Advances in Information Retrieval. ECIR 2009. Lecture Notes in Computer Science, 5478*, 337-349. doi: 10.1007/978-3-642-00958-7_31

Ulrich, K. T., & Eppinger, S. D. (2015). *Product Design and Development* (6th ed.). McGraw-Hill Education.

Wang, Y., Zhou, Z., Jin, S., Liu, D., & Lu, M. (2017). Comparisons and Selections of Features and Classifiers for Short Text Classification. *IOP Conference Series: Materials Science and Engineering*, *261*, 012018. doi: 10.1088/1757-899X/261/1/012018

Waykole, R. N., & Thakare, A. D. (2018). A Review of Feature Extraction Methods for Text Classification. *International Journal of Advance Engineering and Research Development*, *5*(04), 351-354.

Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects-based sentiment analysis. *Expert Systems with Applications*, *39*(11), 10283-10291. doi: 10.1016/j.eswa.2012.02.166

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758-2765. doi: 10.1016/j.eswa.2010.08.066

Zhou, Q., Xia, R., & Zhang, C. (2016). Online Shopping Behavior Study Based on Multi-granularity Opinion Mining: China Versus America. *Cognitive Computation*, *8*(4), 587-602. doi: 10.1007/s12559-016-9384-x