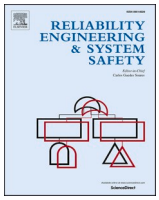
Reliability Engineering and System Safety 216 (2021) 108010

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



Promoting a novel method for warranty claim prediction based on social network data

Sajjad Shokouhyar [a], Sadra Ahmadi [b,*], Mahdi Ashrafzadeh [b]

[a] Management and Accounting Faculty, Department of Industrial and Information Management, Shahid Beheshti University, Tehran, Iran
[b] Cyberspace Research Institute, Shahid Beheshti University, Tehran, Iran

ARTICLE INFO

Keywords:
Warranty claim prediction
Social media analysis
Random Forest method

ABSTRACT

Warranty plays an important role in retaining consumers' loyalty, increasing the competitive advantage and the profit of companies. Moreover, warranty claim prediction based on social media is a novel area, enabling managers to foresee problems in production and take the proper measures to mitigate them. The higher the precision of the warranty claim predictions, the lower the risk the company faces. This paper examines the impacts of utilizing social media data on daily warranty claim prediction. In this paper, we showed that social media data could enhance the accuracy of daily warranty claim predictions. We cooperated with Sam Service Warranty Company that provides warranty and aftersales services for Samsung products in Iran. Warranty operational data along with Twitter data analyses were used to improve the precision of warranty claim prediction. Operational data from Sam Service Company include the total number of warranties, the number of warranties for new customers, and the number of warranties for those who return. A novel framework was presented that uses the Random Forrest algorithm for prediction of the number of daily warranty claims. The results show that our framework improves the accuracy of out-of-sample warranty claims predictions, with respective development at a range of 14.98% to 21.90% across various timeframes. Improving prediction accuracy enables managers to effectively minimize warranty-related costs, inventory levels, waste, and customer dissatisfaction while maximizing the return on investment, profit, efficiency, and customer satisfaction.

1. Introduction

A warranty can be defined as a contractual obligation that manufacturers (or vendors or sellers) incur concerning sale of products [1]. Accordingly, the manufacturers are required to deal with the existing challenges or failures which are faced over a specific warranty coverage period [2]. Although a satisfactory warranty term may be attractive for a significant number of customers, obligations related to such a warranty generate additional costs for the manufacturer [3–7]. According to a 2014 survey by Warranty Week, different big companies worldwide may incur 2-5 percent of warranty costs as a percentage of the sale price. Therefore, to effectively manage the warranty-related costs, the manufacturer has to know in advance about the possible warranty claims for their products [8, 5]. To effectively manage the warranty-related costs while maximizing the return on investment and customer satisfaction, in advance prediction of the warranty claims of the products is necessary for the manufacturers [5]. The warranty claims data are lifetime data that are gathered over the process of servicing the products during the period of a warranty, as well as some warranty terms [9]. Management of the warranty claims data/databases is carried out by many dominant manufacturers worldwide. One of the most eminent applications of warranty databases is the detection of invisible reliability problems emerging at design and/or manufacturing phases during modeling, and the warranty claim data analysis.

Analyzing the warranty claim data enables manufacturers to make better decisions on their financial strategy plans and their warranty policy along with better product designs, which helps them achieve a higher level of customer satisfaction. [8, 2, 4, 5]. By anticipating warranty claims, companies can minimize the inventory level, reduce the costs of the product development process, allow advanced product designing and manufacturing, improve quality and reliability, acquire a competitive advantage in highly competitive markets, and develop financial strategy plans [8, 2, 4, 5]. Warranty data analysis is used as a strategic decision-making tool in supply chain analysis, and it involves selecting/switching suppliers, increasing/decreasing the warranty period/scope, and changing warranty contract terms [2, 4].

* Corresponding author.
E-mail address: Sa_ahmadi@sbu.ac.ir (S. Ahmadi).

https://doi.org/10.1016/j.ress.2021.108010
Received 20 February 2021; Received in revised form 5 August 2021; Accepted 22 August 2021
Available online 4 September 2021

These days, a lot of manufacturers, particularly manufacturers of complicated products, consider warranty as an effective means of boosting sales, building brand loyalty, and gaining a higher level of customer satisfaction [10]. Most researchers use companies' operating warranty data to predict consumers' claims. It should be noted that analyzing the product lifetime data requires high-quality data; however, the quality of companies' operational warranty data is usually very low [9, 5]. For instance, some common limitations and inaccuracies in warranty data include "not being able to determine the exact age of a product", "delays in analyzing data", "delays in customer reporting", and "lack of clarity with regard to the consumer's behavior when using a product". Thus, companies' operational data have inconsistencies in terms of the age and the time of reporting, along with the probability of data censorship, missing data, and data ambiguity. as follows [4, 9]:

1 Inability in determining the precise product age: Product age can be described as the period between the sale and return of the product to the retailer/manufacturer. Warranty claim data are only collected as claims in different groups, and they do not account for the age of the products. This means that the analyst may be unable to achieve the accurate age of the product under warranty, and he/she only knows that the age is in a specific range.
2 Delays in reporting: This is defined as the interval between the time an incident or a failure occurs in a product and the time it is reported, which might be caused by the customer or the company:
   a Company delay: The customer reports an incident promptly; however, it significant time may be needed to enter the report into the warranty claims database to perform analyses. In other words, this leads to a delay between the incident reporting time and the data insertion time for the warrant data analysis.
   b Customer delay: This refers to the case where the customer does not report an incident immediately. Hence, the consumer delay can be defined as the time between the occurrence of an accident and the time the incident is reported.
3 User Behavior: Analyzing warranty data and the factors that cause a product to be returned as warranty claim is an interesting topic for warranty analysts. It is often difficult to identify these factors, which can be due to the user's behavior. For instance, a user may accidentally spill some water on an electronic device.

As social media have penetrated into every aspect of people's lives, customers discuss companies' products and services on social media platforms, including Twitter and Facebook. They express their feelings, preferences, interests, opinions, and problems with products and services. Different organizations which have recognized this are starting their work on social media to help customers in solving their problems. Using social media in warranty analysis can address the three above-mentioned challenges.as follows: [11–14].

1 Product age determination: since consumers on social media mostly mention the purchasing time of a product as they want to know if the bought product still has warranty coverage. Thus, the product's age can be extracted by the manufacturer.
2 Just-in-time reporting: nowadays, social media is simply accessible everywhere, and users usually tweet about the problem they are faced with immediately. This will result in having a lower delay in receiving consumer's feedback.
3 User behavioral details: using social media, consumers usually report to some extend detailed information about the causes of their product failure in their tweets.

The main challenge researchers face regarding social media data involves the high-dimensional data problem (i.e., the number of variables largely associates with the number of observations). In addition, conventional prediction methodologies, including "Linear Regression" and "time series models" usually have poor performance on high-dimensional data.

The use of a machine learning-based framework aimed at overcoming this problem. [15] show that machine learning models have the conditions required for solving these types of problems. This study also provides empirical evidence to show what kinds of models can effectively be used for analyzing social media data. In this study, we add/-compare social media warranty data into/with the companies' warranty data until the results of warranty claim predictions become more rigorous. Combining these two sets of data will help the company achieve more precise predictions in relation to warranty claims, which results in mitigating warranty costs and increasing the company's profit.

To show the effectiveness of our approach, we conduct a comparison between the analysis results on the two following sets of data for prediction of the number of daily warranty claims using the same machine learning algorithm:

1 base model: A basic prediction using only the companies' operational data.
2 social media model: A combination of social media data and the operational data helps to achieve a higher level of precision.

Quantifying the value of social media data in the improvement of warranty claim predictions is possible by comparing the out-of-sample prediction accuracy with the use of these two types of predictions.

This paper improves our understanding of how social media data can be used in the context of managing warranty claim services, which other researchers barely discuss. The innovations of the article can be elaborated in more detail. This paper has three main contributions. The first contribution of this paper is using social media warranty claim data to achieve more precision in predicting warranty claims. Second, we present a new method for a predict warranty claim that combined social media warranty claims data with operational data. Combining these two sets of data will help the company to achieve more precise warranty claims predictions. The accuracy improvement of our social media prediction model is significant across different forecast horizons. Third, this research makes an important contribution to the community of warranty services management. Using this method of analysis, managers will gain broader insights into their decision-about the warranty management.

2. Research questions and Methodology

The present paper seeks to address the research questions raised as follows:

RQ1: How does the warranty claim prediction help managers have more control over supply chain management?
RQ2: Do social media data and consumers' opinions, combined with operational warranty data, help the manufacturer to improve the precision of the warranty claim prediction?
RQ3: To what extent do social media data improve warranty management?

To answer these questions, the following objectives were set:

1 A literature review on warranty claim prediction is performed to investigate the benefits that managers can gain by putting more effort into collecting more data from the consumers.
2 systematic research is performed on warranty claim prediction to propose a new insight for the strategic decision-maker of the companies.
3 A methodological framework is proposed to help managers have a broader insight into the investigation of cause and effect of social media data in warranty claim prediction.

This study attempts to answer the above-mentioned questions

through a systematic approach, as represented in Figure 1.

This systematic approach consists of three steps. In Phase 1, we reviewed the relevant literature to identify the causes of the low quality of the operational warranty data. The results of the review show that operational data suffer from a number of limitations, including "not being able to determine the exact age of a product", "delays in analyzing the data", "delays in customer reporting", and "lack of clarity with regard to the consumer's behavior in using a product". Then, we finalized a list of these causes. In Phase 2, we collected data from Sam Service Company and from the official Twitter page of Samsung. Afterward, we studied the statistics of the operational data and social media data, and a statistical summary of these data is presented in Tables 5 and 7. Finally, we constructed two prediction models, i.e., the social media prediction model and the base prediction model. In the last phase, we first present the findings and managerial implications, followed by discussing the conclusions and future research directions.

3. Literature Review

In today's market, the use of warranty is considered for many purposes, including customer satisfaction, marketing strategy, and legislation. Consumers expect manufacturers not to leave them in case of a failure in their products [1, 16].

Most companies use their operational warranty claim data to analyze and predict future claims. Since most consumers are comfortable with expressing their feelings and putting their problems and questions on social media, we used social media as a rich source of data that can increase the accuracy of predicting warranty claims. Firstly, we discussed the concepts of warranty and guarantee. Then, previous research focusing on warranty and social media was reviewed, and finally, we reviewed the most eminent research on warranty claim prediction using operational data.

3.1. Concepts of Warranty and Guarantee

Warranty and guarantee services can be used as strategic weapons during the constant battle for differentiation and attainment of competitive advantages in service industries. In addition, the possible advantages of guarantees or warranty services are maintenance through reduction of consumers' risk, improvement of service quality, and establishment of service standards ([17, 18]; X. [19]). However, there are differences between these two concepts.

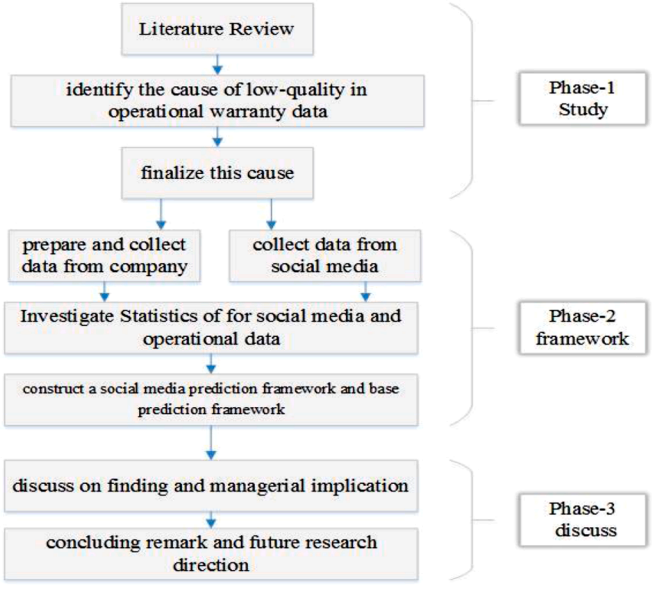


Figure 1. Research Methodology

Under the policy of buyback warranty, the customers have the chance of returning the manufactured products over the period of warranty and obtaining a refund of the sale price from the manufacturers. The unconditional (money-back guarantee) or conditional refunds on predetermined events, e.g., when the number of failures during the period of warranty is more than a specific limit, are taken into account [20]. The replacement or repair of the products which have failures will be free of charge for the customers over the period of warranty under the money-back guarantee policy. If the product fails more than a predetermine number, k, over the period of warranty, the customer can return the product and receive 100% of the money back [17].

The money-back guarantee (MBG) service is typically a real-world program. Money-back guarantee (MBG) is offered by many retailers [17, 21]. From the consumer's perspective, guarantee services are considerably satisfactory in the case that the retailers offer the MBG plans, allowing the return of the products and receiving a full refund. Obviously, the chance of returning the purchased products with such a guarantee attracts a greater number of customers and stimulates demand. Particularly, if the consumers are not capable of judging the quality of the products before their purchase (such as e-commerce), the MBG service increases their confidence, decreases risk [22, 21], and signals the quality of the guaranteed products [23].

3.2. Social media and warranty claim

Today, social media platforms have wide applications as an instrument in businesses for prediction, education, customer engagement, recruitment, decision-making, promotion, advertising, and so on ([24, 15, 12]; X. [25]). Social media platforms enable companies to gain consumers' attention and increase interaction with them to attract more engagement [26–29]. Before making any decisions to buy a product, most customers refer to the user-generated comments on products on the social media for evaluation of the products and services ([12]; L. [30]).

Since researchers have more access to the social media data through APIs, different research works try to indicate the value of such data in supply chains [31–33], green supply chains (Shanshan [34–36]), recycling [37, 38], and restoring [39].

Despite the attention researchers have recently devoted to social media, to the best of our knowledge, no studies have yet discussed the opportunities that social media platforms put in front of managers to make warranty claim predictions. [40] express that from the consumer's perspective on social media, one of the most important questions about products was the warranty. [41] state that customers whose products have problems usually worry about whether the warranty will cover repair fees. Therefore, these authors started working on social media platforms, including Twitter and Facebook, and then they suggested using social media data for analyzing warranty services. [42] suggest that researchers may explore various warranty services through social media. [43] recommended identifying and collecting warranty data from Twitter to perform warranty services, such as warranty claim prediction, eliciting customer feedback on products, advertisement, and so on. Utilizing the operational value of the social media data, our study addressed using social media data in warranty claims prediction.

Consumers can express their feelings about warranty services through social media channels. Positive and negative feedbacks from consumers indicate customer satisfaction and dissatisfaction with the warranty products and services, which can affect other consumers to use the warranty in similar cases. A vast literature is available on the effects of feedbacks from consumers on warranty services. For instance, [44] declares that warranty service providers usually carry out tracking research, in which random selection of some customers is conducted along with close follow-ups. In fact, the main objective of tracking customers is extracting customer sentiments on the warranty services to improve aftersales services of the manufacturer. (X. [45]) assert that the FBNR (failed-but-not-reported) phenomenon shows whether customers

are dissatisfied with the products or the warranty services, which can help manufacturers better understand users' sentiments about the product or the warranty services. [46] show manufacturers can determine customer satisfaction warranty-based feedback, which can help to identify the root causes of defects more systematically. Therefore, we analyzed the sentiments in comments of users through natural language processing (NLP) methods.

3.3. Warranty claim prediction

3.3.1. Warranty data analysis categorization

Products return (warranty return) process as a component of reverse logistics [47]. Figure 2 shows the typical lifecycle of failed products. This process begins with product manufacturing and ends with the return of the products to the manufacturer [4].

Several studies have mentioned that the long-term relationship and trust of consumers are important components of supply chain success (Suhong [48]). In contrast, insufficient attention to this relationship and the lack of mutual trust often lead to the failure of supply chain partnerships [49]. Therefore, companies nurture long-term relationships with customers in the supply chain by providing warranty services for their products.

Most companies offer product warranty to guarantee product quality, and to gain consumers' trust in their products. Good warranty service is an important factor in increasing customer commitment [50]. warranty services increase the demand from consumers; however, it puts an additional burden on the manufacturers in the form of warranty costs [4]. As a result, many researchers have developed some frameworks for warranty data analysis to help manufacturers reduce product failure rates and warranty costs.

The warranty data analysis is a process that extracts useful information and helps the decision-making process using statistical or computer algorithms (e.g., random forest models, machine learning models, and so on) [2]. Warranty data analysis is categorized into five areas, as shown in Figure 3 [2, 4].

In this study, we focused on warranty claim prediction because

1 As far as we know, not much work has been done in the field of warranty claim prediction to increase prediction accuracy using social media.
2 Accurate prediction of the warranty claims is vitally important for managerial decision-making and financial plans [4].
3 From a manufacturer's perspective, warranty claim prediction relates to managing a warranty policy; thus, warranty claim prediction is a critical dimension of a warranty program management. To effectively manage a warranty program, managers essentially plan the warranty policy of the manufacturer based on warranty claims prediction [8, 2].
4 Having accurate warranty claim prediction can improve aftersales services and customer satisfaction [8].

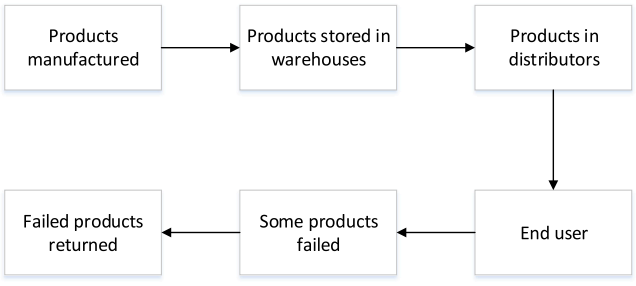


Figure 2. The Typical Lifecycle of a Failed Product

3.3.2. Warranty claim prediction methods

Warranty claim prediction is a prediction process that predicts the expected number of consumer claims in the duration of the warranty. It should be noted that warranty claims incur enormous costs for the manufacturers, at a typical range of 1% - 5% of the total sale price [51]. The cost of the warranty is almost proportionate to the demand for the product repair. Furthermore, the number of demands for repairing is a major factor in determining the size of the aftersales services department [1]. Therefore, accurate prediction of the total number of repair demands (i.e., warranty claims) over a specific period is very important for the manufacturers.

In the field of predicting warranty claims, many researchers have proposed a number of frameworks, which use different prediction techniques. Using different methods, most researchers have designed an accurate model to predict the number of warranty claims [1, 16]. These methods are shown in Figure 4. [8, 5]

3.3.2.1. Warranty claim prediction methods in lifetime distributions and stochastic procedure. There are three different methods to predict the number of warranty claims, including lifetime distributions, stochastic process and machine learning, which are indicated in Figure 4. The data used in lifetime distributions were related to the product life data (B. K. [52, 53, 16]). We use two sets of data to predict the number of warranty claims. The first set consists of operational warranty data, including the number of warranties by total, returned customers, and new customers. The second set of our warranty data is related to social media, which includes comments and posts. According to the difference between the data used in this paper and the lifetime distribution methods, the lifetime distribution methods are not appropriate for our paper. In addition, a stochastic process reflects having a system with relevant observations at specific times and a random variable as the outcome, which is the observed value at every specific time. Therefore, the stochastic process is an unrelated way to build a prediction model in our paper.

3.3.2.2. Warranty claim prediction method in machine learning. We used two sets of data to predict the number of warranty claims. As stated, the second set of our warranty data is related to social media. The main challenge researchers face regarding social media data involves the high-dimensional data problem (i.e., the number of variables largely associates with the number of observations). The use of a machine learning-based framework aimed at overcoming this problem.

Concerning machine learning, various algorithms, such as artificial neural networks, deep learning, Logistic Regression, Support vector machine, Support vector regression, linear regression, decision tree, and random forests, have been employed for the prediction of warranty claims according to data of operational warranty. (B. [54]) proposed employing a specific type of artificial neural network, i.e., the radial basis function (RBF) networks, to conduct the prediction tasks. They used a neural network model for prediction of year-end warranty performance when the 'maturing data' phenomenon was present. [55] applied an MLP (multi-layer perceptron) for prediction of the warranty costs according to single vehicle variables (including age, mileage rate per month, as well as road conditions index), and the total manufacturing quality fluctuation risk (various technical groups). [56] presented a methodology for the optimization of the FRW/PRW (free replacement warranty/ pro-rata warranty) policy based on the anticipated warranty expenses for the manufacturers. They demonstrated a combined free replacement and pro-rata warranty policy (in the case of the item failure prior to the end of the its warranty period, it is substituted or repaired as per usual; however, only for the amount based on the costs which depend on the age of the item when it fails.), which was analyzed as the warranty model for a type of light bulbs. A neural network model was employed for the prediction of the light bulb reliability features according to the data obtained by testing light bulbs in different operational situations. (J. [57]) utilized warranty data for

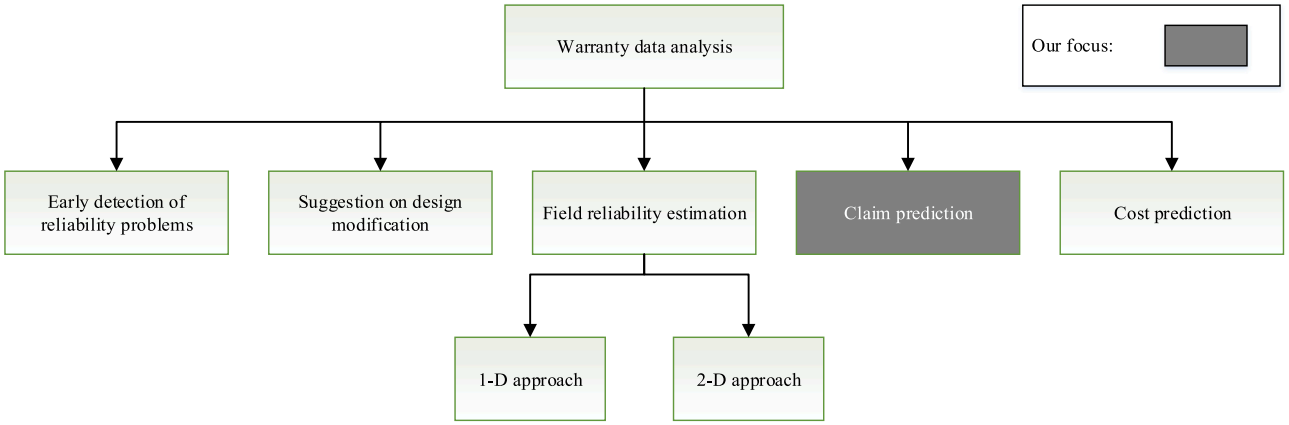


Figure 3. Categorization of Warranty Data Analysis

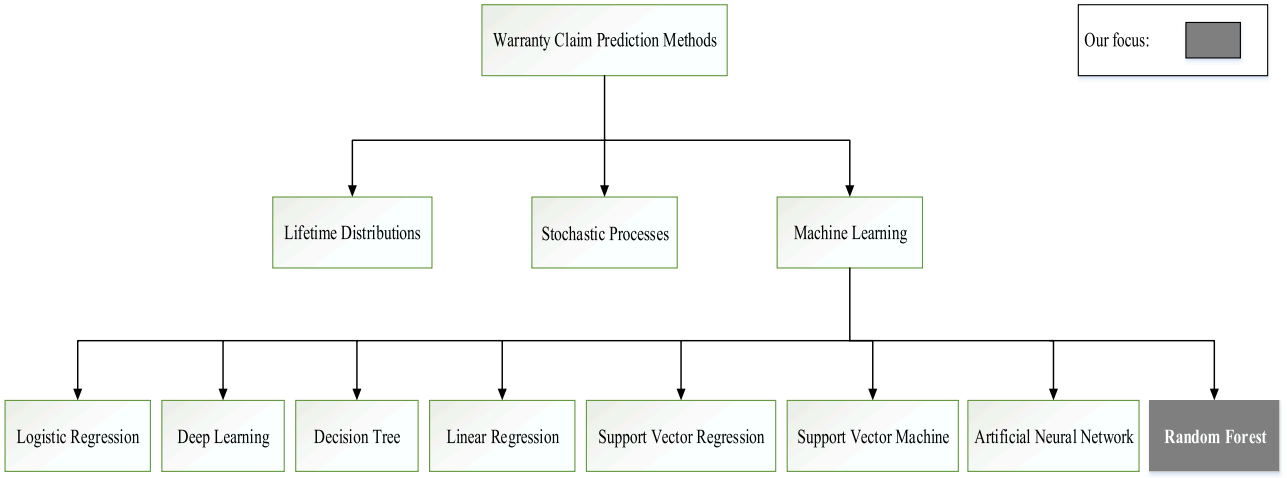


Figure 4. Warranty claim prediction methods

optimization of the spare parts inventory with subsequent improvement of the level of aftersales services, while controlling the costs. They constructed an SVM-ABC (support vector machine-ABC classification) model to classify the spare parts. They studied the reliability features of the parts in the warranty data, taking into account the reliability features parameters of the spare parts in usage (Mean Time between Failure), supply features (replenishment lead time as well as supplier scarcity), part costs, and part criticality. [58] given a technique for adaptation of historical maintenance warranty data, while comparing the Artificial Neural Network (ANN) and multiple linear regression approaches. Their results showed that the best results were obtained for both MSE (mean-squared error) and R^2 (correlation coefficient) when all independent variables were combined. [59] addressed the problem of products' terminal call rate (TCR refers to the information on the amount of funds to be reserved for product repairs during the warranty period) prediction during the warranty period. They developed a series of deep learning models on a data set obtained from a manufacturer of home appliances. [43] they said, warranty claims data, which are typically stored in warranty databases, may be analyzed to improve quality and reliability and reduce costs in areas. Hence, there are three challenges, first, to accurately recognize production faults from these multiple sources of heterogeneous textual data. Second, accurately mapping the identified production faults with the appropriate design information and third, using these mappings to simultaneously optimize costs, design parameters and tolerances. Therefore, they suggested a data mining based method using Self Organizing Maps (SOM) to draw information from the warranty database and to relate it to the manufacturing data. [60] introduced a machine learning method for warranty claim prediction by integrating available information sources such as Logged Vehicle Data and Warranty Claim. The experimental results obtained from a large data set of heavy duty trucks. [61] illustrated a method for prediction of warranty repair claims for automotive units according to joint on-board diagnostic together with historical warranty repair data. They also evaluated the performance of Support Vector Machines, Random Forests, and Decision Trees on a specific dataset. It is possible to integrate large volumes of data logged over the operation of a vehicle, such as trends and emerging patterns that caused quality problems in vehicles, to provide the model of utilization patterns for primary predictions. [62] offered a new method towards the early detection of quality issues using linear regression and logistic regression to predict the failures of a given component across the large population of units. They used the sensor data and the operational warranty data.

3.3.2.2.1. Some limitations of machine learning methods. To the best of our knowledge, no study has used social media data and analysis in warranty claims. We link the customers' opinions on social media to warranty claim prediction. In a nutshell, we collect customer comments and posts containing problems with products from Twitter, then analyze emotions in them. Afterward, machine learning models are employed for the prediction of the number of warranty claims. We use a machine learning model (random forest model) for the prediction of the number of warranty claims because the random forest algorithm has the best performance on operational warranty data and social media data [15,

63,61,85]]. In addition, the random forest model performs well in small samples with high dimensionality [15]. Therefore, we do not consider other machine learning models because they have some limitations, which we have presented in Table 1.

4. Methodology

In this section, the framework to predict warranty claims will be outlined. This framework consists of four steps, i.e., data collection, data preparation, the prediction model, and the evaluation of the results. Figure 5 summarizes the framework proposed for predicting warranty claims.

4.1. Step1: Data Collection

We collected data from the company's Twitter page, and then we obtained the company's operational data. There are various sources of gathering customers' reviews of products, such as social media platforms (Twitter, Facebook, Instagram, Tumblr, Skype, WhatsApp, etc.), e-commerce platforms (Amazon, Google My Business, Yelp, etc.), customer search data (Google Trends), online ads, E-mail Marketing, product surveys, crowd funding platforms(Kickstarter) and so on. We summarized all the limitations related to these resources in Table 2 ([26]; D. [64, 15, 65–67])

We collected data from Twitter. The advantages of customers' reviews of products on Twitter are:

1 Twitter is one of the most powerful social media platforms that has gained popularity among social media users in the world and analysts. According to the statistics released in the first quarter of 2019, Twitter has 330 million active users ([68,86])
2 Twitter contains an abundance of data. Unlike other social platforms, more user data (tweets) are publicly available. This lends Twitter to be an excellent resource for data analysis. [40]
3 There are three different ways to access Twitter data: Twitter Search API, Twitter Streaming API, and Twitter Firehose. [40]
   a Twitter Search API provides access to a database of previously posted tweets. Through the Search API, users request tweets that match some sort of "search" criteria. The search criteria can be keywords, usernames, locations, places, names, etc. The maximum number of tweets that can be reached is 3,200, regardless of the search criteria for an individual.
   b Unlike Twitter's search API, Twitter's Streaming API searches tweets as they are posting in real-time. With Twitter's Streaming API, search criteria (keywords, usernames, locations, places names, etc.) are initially set, and the tweets that match the criteria are provided to the user. Studies have estimated that users can expect to receive approximately 1% of the tweets using Twitter's Streaming API. One way to access 100% of tweets in real-time is through using Twitter "Firehose".
   c Twitter Firehose is handled by two data providers, GNIP and Data Sift. The difference between Twitter's Streaming API and Twitter Firehose is that Twitter Firehose provides access to 100% of the posted tweets; however, unlike Twitter's Streaming API, it is not available for free. On the other hand, Twitter Streaming API is available for free.
4 All Twitter APIs provide encoded data in JSON format. Each tweet is associated with an author, a message, a unique ID, a timestamp showing the time it was posted, and sometimes geo metadata is provided by the user. [40]

Table 1
Some limitations of machine learning methods in dealing with our work

Machine learning	Limitations	How Random Rorest Address These Issues
ANN	Poor performance in small samples with high dimensionality data set. [15]	[15] shows empirically, the random forest model has the best performance in small samples with high dimensionality data set.
Deep Learning	It is difficult to train and contained a large number of hyperparameters. [81]; it needs large datasets to train the models. [59]	The random forest model does not require much effort in tuning hyper-parameters [63]; in addition, Only two parameters of the RF algorithm need to be adjusted for optimal performance. [82]
Logistic Regression, Linear Regression, SVR and SVM	Have not performed well on social media and operational warranty data. ([15]; K. [83, 84, 61])	The random forest model has performed well on social media and operational warranty data. [15, 63, 61]

The methodology used to collect data is illustrated in the present section, consisting of the operational data and the social media data. We discussed how to collect and combine the data from Sam Service Company and Twitter.

4.1.1. Operational Data

We collaborated with Samsung Company, which provides aftersales services for its products, such as smart phones, TV and audio devices, home appliances, smart devices, and computing. To evaluate the value of the social media data, we collected and compared two different datasets, i.e., internal operational data from Sam Service Company (Samsung's representative for aftersales services in Iran) and publicly available social media data from Twitter page of Samsung Company (official page). We used the operational data to provide comparisons of the accuracy of the social media model and the base model. The internal operational data, obtained from Sam Service Company, consist of raw warranty data.

Raw warranty data include total number of daily warranty claims, daily warranty claims for new customers (who haven't previously used a warranty), and daily warranty claims for returning customers (who recently used a warranty). We collected detailed data on the marketing activities of the company, which is internally available for the firm to determine the value of social media data combined with the data related to marketing efforts. As a result, inclusion of data obtained from the email campaigns in our base prediction makes independent consideration of the effects of marketing interventions from the social media data possible. A principal daily variable associated with the company's marketing efforts was constructed. We documented the company's warranty data for the period between 2018-01-01 to 2018-12-31.

4.1.2. Social network data

Twitter is an online social network service, which provides users with the chance of sending short 280-character messages, known as tweets. It can be regarded as the most popular social media platform, where people express their feelings and opinions, and it is one of the largest social media platforms where companies engage in promoting their products and informing and educating their customers, stay in touch with their audience, and so on. Over the last quarter of 2019, Twitter was reported to have 152 million monetizable daily active users (MDAU) across the world [87]. One of the most significant advantages of using Twitter is the possibility of accessing tweets through a public Application Programming Interface (API). For this purpose, a Python program was developed to extract all posts on Twitter through the API. Because of the accessibility of an API that can store tweets accessible by researchers and the favorable characteristics, including filtering based on variables such as location and keywords, scholars and research communities are interested to investigate the capabilities of Twitter further than a social media. As mentioned earlier, besides the company's operational data, social media data relevant to Samsung in Twitter were also collected. Samsung Company has an official page on Twitter with

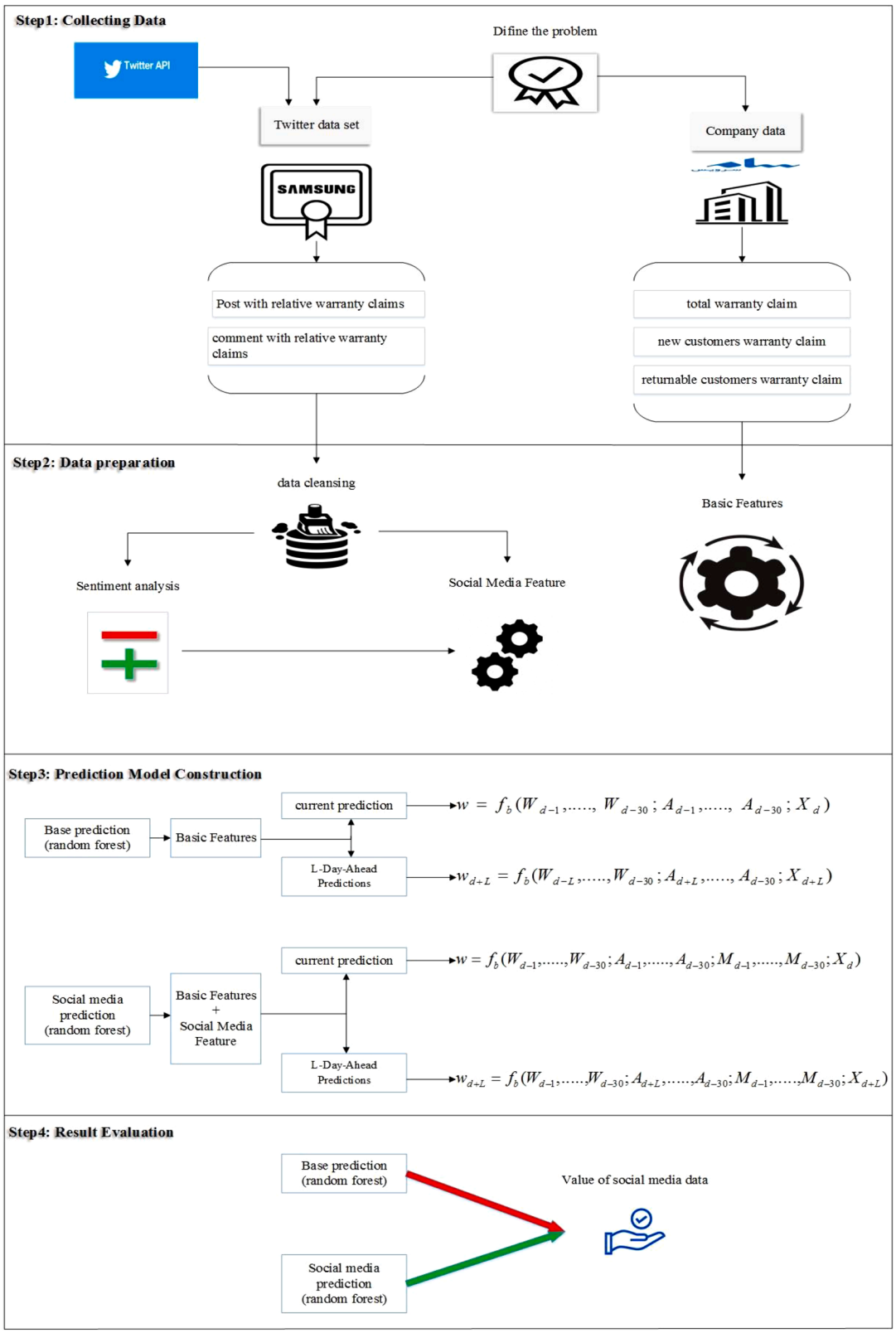


Figure 5. The Prediction Framework

more than 11 m followers. Interaction with a company's Twitter page is also possible for the users in four ways. Initially, writing remarks is possible with the use of the "comment" option under every post. Moreover, a "like" button makes endorsement or approval of a post possible. Furthermore, a "retweet" button makes publishing the company's post possible. Finally, they can use "quote" to publish a company's or a user's post. If a user "retweets", "likes", "comments" and "quotes", her friends can probably observe the same post in their news feed.

Sentiment analysis (or opinion mining) mainly focuses on the extraction of subjective information (e.g., emotions, opinions) from

Table 2
Summary limitation of sources of customers' reviews

sources	limitations
Social media platforms	Numerous sources
E-commerce platforms	accessing the source of data for analysis, unrelated to warranty data
Customer search data	not obvious how to choose search terms, needs a level of trust in the data
Online ads	Potentially disturbing for customers
E-mail Marketing	Undelivered E-mail, E-mail response decay, Render ability, Expenses
Product surveys	Potentially disturbing for customers

tweets. We adopted the methods of natural language processing to achieve the sentiment in the comments of users. Based on an influential work on natural language processing [69], a recursive neural tensor network (RNTN) was employed. This method outperforms all previous methods on several metrics, such as negative classification accuracy, somewhat negative classification accuracy, neutral classification accuracy, positive classification accuracy, and somewhat positive classification accuracy. We classify the message according to its polarity, i.e., whether it is positive, negative, or neutral. Aggregation of emotions into total opinions in one day was considered to measure daily emotions.

4.1.2.1. Biases in dealing with social network. The essential drawback of Twitter is the lack of guidelines on what data users receive and how much; this causes one to doubt whether the data being sampled is a true reflection of actual Twitter behavior. We should consider variables such as data collection topics (keywords and hashtags), language, geo-location, data source, etc, as they affect Twitter performance. The following are some of the most important biases related to this research. [70]

I Keywords and hashtags: One way to address this bias might be to create more specific parameter sets with different keywords; in this study, the metrics to extract the tweets are taken from three various sources. Such as:
  a A comprehensive literature review of the warranty policy
  b Warranty policy Terms on companies' website
  c Evaluation of the keywords with the help of some experts in the warranty
II The source of data: We consider another factor, "the source of data", which acts as reasonable proxies in the context of microblogs. Source credibility means the information comes from a credible source.

4.2. Step2: Data Preparation

In a prediction model using machine learning, raw data cannot be used directly. This stage in the prediction modeling is known as data preparation which refers to the technique of preparing data, such as data cleaning (identification and correction of mistakes, errors, etc. in the data), data transformation (modifications in the scale or distribution of the variables), feature engineering (extracting new variables from the existing data), feature selection (identification of input variables with the highest relevance to the task), and so on. Therefore, we first cleaned the data using the cleaning method and then selected the variables that should be applied as inputs and outputs in a prediction model.

4.2.1. Data cleaning method and the corresponding limitations

We collected Sam Service warranty data, which is providing after-sales services in Iran. For each comment and post extracted from Twitter, some useful attributes have been provided, including date, time, and the location of the tweets. These attributes enable us to classify the tweets and to analyze them based on our desired location. Therefore, the tweets have been filtered out by the user geo-location attribute.

It is possible to manually insert the location shown in the user profile, so the user is able to enter any location he/she wishes. Therefore, we excluded the tweets with the following features from the final set of tweets: (1) the location variable was considered blank; (2) the location variable possessed an unknown location or a random word, e.g., 'would not you like to know', 'somewhere over the rainbow', and 'here'; and (3) the location variable did not have enough specificity, e.g., 'earth', 'global'.

4.3. Step3: Predicting Framework and Machine Learning Model

In this section, we compared two predictions, i.e., a "base prediction" which merely consists of operational data as input variables, and a more accurate prediction, called "social media prediction" which comprises both operational data and Twitter data. Operational data were collected from Sam Service Company, while social media data were collected from the official Twitter page of Samsung. The neural network models are not considered due to their poor performance small samples with high dimensionality because these models have considerable flexibility, which makes them convenient to over-fit [15]. The same machine--learning model was fitted for the two prediction models. To measure the improvement obtained by using Twitter data, the difference between accuracy values was calculated. In addition, we used cross-validation to identify the best number of variables. This section is divided into two main parts:

1 Section 4.3.1 brings a discussion on the prediction settings and the total framework employed for the construction, training, and testing of the prediction models.
2 Section 4.3.2 presents a description of the implementation and the results obtained from the machine learning method.

4.3.1. Base Prediction Model

In the Base Model, the warranty claim on day d is supposed to as a function of the warranty claim data (products returned to use warranty) in the previous month (previous thirty days) and x_d was considered as the features related to day d:

$$w = f_b(w_{d_{-1}} \ldots . w_{d-30}; A_d \ldots . A_{d-30}; x_d) \quad (1)$$

In which χ_d indicates the features related to day d, including the day of the month; W_d denotes the number of total warranties; and A_d represents the advertising variables, such as the number of advertisements on day d. This mechanism has been used in the literature for daily predictions based on social media data [15].

4.3.2. Social Media Prediction Model

In the "social media prediction" model that incorporated social media data into the operational data, we assumed the same structure for variable notations as those used in the base prediction model. The number of the current warranty claims is a function of the number of the past warranty claims data and social media data during the previous month, and x_d is the characteristic related to that day:

$$w = f_s(w_{d-1} \ldots . w_{d-30}; A_d \ldots . A_{d-30}; M_{d-1} \ldots . M_{d-30}; x_d) \quad (2)$$

In which M_d indicates the social media characteristics on day d, including the number of comments, the number of words in the comments, and the sentiment of the average comments. According to the specifications, W_d may indicate the overall warranty. To estimate and compare the described models, we divided the data into training and testing sets. The in-sample training set includes 240 days $(D = 240)$ and the out-of-sample testing set includes 120 days$(T = 120)$. The dataset is classified into the training set $1, \ldots, D$ and the validation set $D + 1, D + 2 \ldots, D + T$. A cross-validation procedure is employed to select the hyper-parameters of the proposed models, including the number of characteristics to consider in the random forest model. For every possible value

of the hyper-parameter, the performance was evaluated with the use of the ten-fold cross-validation with three repetitions [15, 71]. Random partitioning of the training set takes place into ten subsets of equal size. Of the ten subsets, one subset is maintained as the validation dataset to test the model's performance, while the other nine subsets can be employed as training data. Ten repetitions (the folds) are considered for the process, while every subset is employed just once as the validation data. Next, the performance of the model on every subset is averaged. The process is repeated another three times to average out errors. As a result, the performance for each potential value of the hyper-parameter was measured, and then the value of the hyper-parameter providing the best performance was chosen.

4.3.3. L-Day-Ahead Predictions

The dataset $\{w_{1,D}; A_{1,D}; M_{1,D}; X_{1,D}\}$ was employed for $f_b(1,D)$ as well as $f_s(1,D)$, i.e., the base and social-media trained models, respectively, with the use of the training period (periods 1 - D) data. When the prediction model is made for the out-of-sample test period, the model parameters and hyper-parameters are re-trained (re-estimated) every day. We construct the prediction warranty claim for day D +1 using all previous data until day D, in addition to the advertising data on day D + 1, to re-train the model. According to the updated model, the data from the last thirty days and future advertisement plans were used as the model input to get the prediction. Accordingly, for prediction of warranty claims at day D + 1, the model estimated through data from day 1 until day N, i.e., $f_b(1,D)$ and $f_s(1,D)$ was used. In the case of predicting the warranty claim at day $D+2$, the model is re-estimated with the use of data from day 1 until day $D+1$, while the updated model is fitted with updated variables through incorporation of the data on day D + 1. Such an updating mechanism for the out-of-sample prediction is widely employed in previous studies [15, 72]. In the above discussion, we predicted warranty claims for the next day. From the practical perspective, companies have to predict farther ahead to adjust their operational decisions. As a result, the L-day-ahead prediction is also constructed. It was estimated how historical information until day d affected the warranty on day d+ L. Besides, it was determined if enhancement of the precision of next-day predictions could have the required robustness to employment of longer lead times, at a range of one to thirty days.

$$w_{d+L} = f_b(w_{d-1}.....w_{d-30}; A_{d+L}.....A_{d-30}; x_{d+L}) \tag{3}$$

$$w_{d+L} = f_s(w_{d-1}.....w_{d-30}; A_{d+L}.....A_{d-30}; M_{d-1}.....M_{d-30}; x_{d+L}) \tag{4}$$

4.3.4. Machine Learning Model

The approach described in Section 4.1 is general and applicable in any prediction model. The machine-learning model is applied in the current section of the study. Machine-learning models are used because of the problem's high dimensionality [73, 74]. The number of independent variables is significantly higher than the sample size, which results in high dimensionality. Simple linear models, including moving-average and autoregressive models, cannot produce identifiable estimations, however machine-learning models are adept at dealing with high-dimensional problems. [15]. Another critical reason for applying machine-learning model is that social media data and even previous warranty data possibly do not influence the present warranty claims linearly.

We implemented one machine-learning model, i.e., the random-forest model. In the Appendix, we presented a pseudo-code of Random Forest, and Figures 9 and 10 provide a specific example of Regression Tree and Random Forest. The most significant benefit of the random forest method is its capability for generation of results with high accuracy, which has led to its popularity as a statistical learning method from the practical perspective. Furthermore, some studies have shown that this algorithm works better for social media data than other algorithms. For example, [15] have shown that the output of the Random Forrest algorithm is much more accurate than the linear regression algorithms, such as Lasso regression, Forward selection, SVM with line kernel, SVM with radial kernel, and Gradient boosting model (GBM). [63] have shown that random forest regression achieves high performance on different tasks for highly structured data, such as social information. [63] have shown that a random forest (RF) model does not require much effort in tuning hyper-parameters and that it performs effectively.

4.4. Step4: Result Evaluation

Different evaluation metrics are used for model performance evaluation, and comparison in regression problem such as Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Normalized Root Mean Square Error (NRMSE) etc. Some significant disadvantages of the evaluation metric presented in Table 3. [75–77]

Finally, we compared the outputs of the models with each other. We measured the prediction accuracy using the mean absolute percentage error (MAPE)(see [26, 15] for further examples) and the Mean Absolute Error (MAE) (see [75] for further example) for both the base model and the social media prediction model.

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|W_{t+i} - \widehat{W}_{t+i-l.t+i}\right|}{|W_{t+1}|} * 100 \tag{5}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\left|W_{t+i} - \widehat{W}_{t+i-l.t+i}\right|}{N} \tag{6}$$

W_{t+i} is the actual amount of the warranty for day t + i. In addition, on day t + i − l, the predicted value of the warranty for day t + i is indicated by $\widehat{W}_{t+i-l.t+i}$. N is the number of the test sample. The advantages of these evaluation metrics are presented in table 4. [75, 78, 79]

5. Data and Empirical Setting

For investigating the presence of the association of social media warranty data on Twitter in terms of warranty claim prediction, a random forest algorithm was implemented. As a contribution, we showed that considering social media data would improve the accuracy of the daily warranty claims prediction estimated through the out-of-sample mean absolute percentage error (MAPE).

5.1. Statistics of Operational Data and social media

In this section, we review the statistics of the operational data. In a perfectly symmetrical distribution, the mean and the median are the same. In our sample, the mean and the median are close to each other, indicating that the data have a relatively symmetric distribution. According to the high standard deviation in our sample, the data points were distributed across a broad range of values, and the data points tend to be far from the mean.

Table 5 indicates the summary statistics of the operational data (in

Table 3
Some important disadvantages of the evaluation metric

Evaluation Metrics	Disadvantages
MSE	It is sensitive to the outliers
MAPE	Lacks a statistical theory on which to base itself
MAE	1. It is an absolute measure 2. Emphasizes larger errors than smaller ones
RMSE	Very sensitive to the attendance of inaccurate data when compared to MAE
NRMSE	similar to the disadvantages of RMSE

Table 4
Some important advantages of the MAE and the MAPE

Evaluation Metrics	Advantages
MAPE	1 The MAPE is generally used for evaluating the robustness 2 The MAPE is a very intuitive explanation in terms of relative error. 3 The MAPE is the percentage error that more significant for decision-makers to pursue the right decision. 4 Provides an easy metric of judging the matter of errors
MAE	1 Its meaning is more intuitive because it is a linear measure 2 The average size of prediction errors when negative signs are ignored 3 allows us to examine the size of our predicted errors

Table 5
Summary Statistics of Operational Data

Variable	Mean	Median	Std. Dev	Max	Min
total warranty claim	776.42	784.5	138.40	996.00	524.00
returnable customers warranty claim	121.81	119.00	43.30	199.00	51.00
new customers warranty claim	492.33	463.00	126.46	832.00	259.00
AD	2.79	2.50	2.62	0	10

the Appendix, Table 6 presented a summary set of the operational data). On average, there are 776.42 total warranty claims per week, 121.81 returned customer warranty claims, and 492.33 new customer warranty claims per week. Warranty claims from new customers were responsible for about 63.41 percent of the overall warranty claims. The company receives a maximum of 993.00 total warranty claims per week in our sample. On average, there are 2.79 advertisements per week.

Figure 6 shows the warranty claim rates of Samsung mobile phone products (collected from Sam Service Company).also the features of the warranty claims data are as follows:

1 Aggregation of the warranty claims is performed monthly.
2 Long-term warranty plans are considered to sell the products.
3 The products are repairable. Testing or repairing the received (or claimed) products is carried out.

In this section, we review the statistics of the social media data. Table 7 shows a summary of the statistics for the social media data (in the Appendix, Table 8 presented a summary set of the social media data). Normally, the company had 29.72 posts a week over the period of the study, with every post having an average of 16.51 words and every post generating an average of 39.82 comments by the users. Moreover, the company receives a maximum of 69.00 comments and a minimum of 14.00 comments per week on its page. The median and standard deviation for each comment are 42.00 and 16.78, respectively.

Figure 7 shows the posts by Samsung, the comments about issues with Samsung's mobile phone product, and customer requests for how to use the warranty (collected from Samsung's page on Twitter).

Table 6
Sample dataset for operational warranty claim

Days	Total warranty claim	Repeated costumers of warranty claim	New costumers of warranty claim	Unknown costumers	advertisement
1	86	17	69	0	2
2	92	19	62	11	1
3	57	12	46	9	0
4	99	21	63	15	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
m-3	126	19	78	29	1
m-2	93	29	64	0	2
m-1	116	38	78	0	0
m	121	39	63	19	1

5.2. Discussion and the value of social media data

We implemented a machine-learning model (random forest) to predict warranty claims by having and lacking access to the social media data. Here, we discuss and emphasize the value of social media data combined with the operational data. To do this, the performance of both machine-learning models is compared. Table 9 summarizes the out-of-sample and in-sample MAPE and MAE for random forest models. Out-of-sample MAPE for the social media model is equal to 10.39, and for the base model, it is equal to 12.74; in addition, Out-of-sample MAE for the social media model is equal to 9.64, and for the base model, it is equal to 12.05 indicating an improvement in warranty claims prediction. In this section, the robustness of the results concerning the prediction horizon is first studied. Secondly, we discuss the robustness of the results with respect to the prediction variable (i.e., new and returning warranty) and different times.

5.2.1. Predicting result of the random forest model

As described in Section 4, we used the random forest model to predict warranty claims with and without social media data for total warranty claim data. Now, we can compare the performance between the two prediction models to evaluate the value of the social media data. We present the analysis results to predict total warranty claims one to thirty days ahead. Table 10 and Figure 8 present the accuracy of out-of-sample prediction and the improvement of accuracy for overall warranty. Table 10 presents both of the prediction models based on the prediction lead-time.

Note we test whether the differences in predicting accuracy are statistically significant by performing a t-test of the differences in MAPE. The p-values are expressed as the strength of evidence with little or no evidence $(p > 0.1)$, weak evidence $(0.05 < p \leq 0.1)$, evidence $(0.01 < p \leq 0.05)$, strong evidence $(0.001 < p \leq 0.01)$ and very strong evidence $(p \leq 0.001)$ [80]. We used three symbols for the p-value. The symbols are as follows: *p <0.1; **p <0.05; ***p < 0.01. For example, in Table 10, the L=1 null hypothesis can be rejected at 0.05 level of significant. It can be explained that the two models do not have the same predictive performance.

Table 10 shows that for both models, i.e., the base prediction model and the social media prediction model, the prediction error increases as the prediction lead-time increases. In the case of five days for the lead-time, MAPE of 10.86 % is generated by the prediction with social media data, while the base prediction model results in a higher MAPE of 13.59%, with 20.59% of corresponding improvement in the prediction accuracy as:

$$\frac{\text{Base MAPE} - \text{SocialMedia MAPE}}{\text{Base MAPE}} \times 100 \tag{7}$$

$$\frac{\text{Base MAE} - \text{SocialMedia MAE}}{\text{Base MAE}} \times 100 \tag{8}$$

Figure 8 shows a summary of the improvements obtained by using the statistical learning instruments and incorporation of the social media data. This result shows consistency over various prediction lead times. According to these results, incorporating social media data in predicting warranty claims provides a superior performance in comparison with that of the operational data.

5.2.2. Robustness with respect to new and returning customers

The purpose of this section is to perform a comparison of the performance of the new and returning customers' datasets. Two various

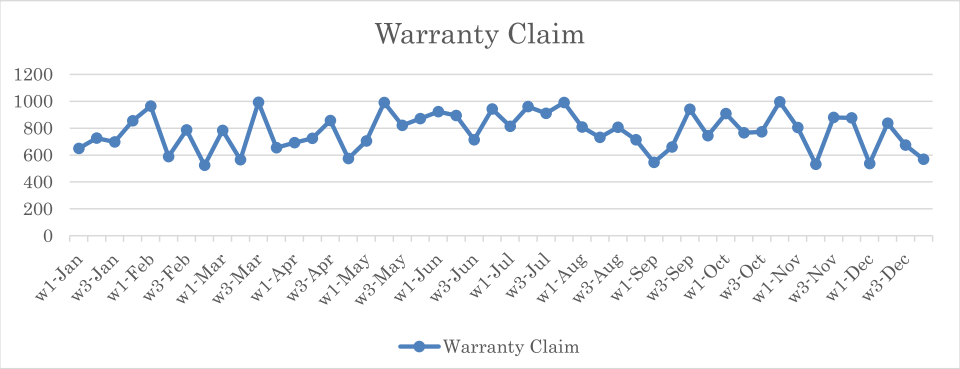


Figure 6. Warranty Claim Rates of the Samsung Mobile Phone in Sam Service Company in the Case Study

Table 7
Summary Statistics for Social Media Features

Variable	Mean	Median	Std. Dev	Max	Min
Number of Comments	39.82	42.00	16.78	69.00	14.00
Number of Posts	29.79	31.5	13.99	56.00	7.00
Average Length of Comments	13.86	12.19	1.76	61	1
Average Sentiment of Comments	2.43	5.47	0.23	7	1

Table 8
Sample dataset for social media warranty claim

days	No. of Posts	No. of Comments
1	4	2
2	5	3
3	2	1
4	3	0
.	.	.
.	.	.
.	.	.
m-3	8	6
m-2	10	9
m-1	6	4
m	11	5

datasets are considered for comparisons. The same analysis for warranty claims made by returning and new customers was performed, and the results were summarized in Tables 11 and 12.

We implemented a data analysis method to predict new customers and returning customers' claims one to thirty days ahead. The values of MAPE, MAE, and relative improvements are shown in Tables 11 and 12. The relative improvements in the results indicate that utilizing social media data leads to a significant improvement in the predictions. The improvements range from 4.53 to 19.16 percent for new customers and from 7.76 to 19.08 percent for returning customers. As stated before, the total warranty claims are divided into two parts, i.e., returning and new customers. The best results were obtained using the social media dataset for new customers with a MAPE range of 9.73 % and 16.39% and MAE range of 5.52 % and 9.81%. For the social media dataset of the returning customers, the obtained MAPE range from 12.94 % to 21.04% and MAE range of 2.32% to 3.70%.

5.2.3. The effect of time on predicting warranty claims

In this section, the effect of time on the perdition of warranty claims is shown. First, the effect of time on the time of collecting training data and testing data are shown. In next section, the effect of time on the perdition models and dataset size are shown.

5.2.3.1. Different time periods. The current section compares the prediction models' performance at different times. The models' performance is compared based on different time predictions. In Tables 13–20, we documented the training and testing sets from 2018-1-1 to 2019-4-30, respectively. Various starting times of the training set were selected, and the analyses were repeated for validation of the robustness of the findings over different times.

Tables 13–20 show the results of the base prediction model and the social media prediction model. In Tables 13–20, we made 1-day, 3-day,

Table 9
Comparing MAPE and MAE for the Statistical Learning Model

Models	MAPE (%)		MAE (Number)	
	RF (in-Sample)	RF (Out-of-Sample)	RF (in-Sample)	RF (Out-of-Sample)
Base prediction	11.62	12.74	10.90	12.05
Social Media prediction	10.06	10.39	9.31	9.64

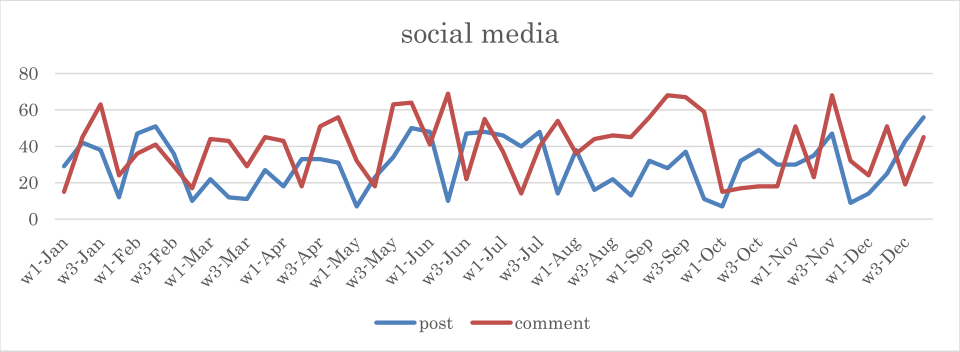


Figure 7. Comment and Post Data of Samsung Mobile Phone Product on Samsung's Page on Twitter

Table 10
Comparing MAPE and MAE for the both Models

Lead Time	MAPE (%) Base	Social Media	Relative Improvement (%)	MAE (Number) Base	Social Media	Relative Improvement (%)
L=1	12.74	10.39	18.45**	12.05	9.64	19.98
L=2	13.38	10.45	21.90***	12.71	9.71	23.62
L=3	13.43	10.91	18.76**	12.76	10.18	20.23
L=4	13.93	11.23	19.38***	13.27	10.50	20.84
L=5	13.59	10.86	20.09***	12.92	9.94	23.07
L=6	14.35	11.77	17.98**	13.70	11.06	19.31
L=7	14.59	12.09	17.14**	13.95	11.39	18.36
L=8	15.37	12.22	20.49***	14.75	11.52	21.88
L=9	15.51	12.87	17.02*	14.89	12.18	18.16
L=10	15.83	12.97	18.07**	15.22	12.29	19.25
L=15	17.02	14.19	16.63**	16.44	13.54	17.64
L=30	20.83	17.71	14.98**	20.34	17.14	15.71

Note: *p <0.1; **p <0.05; ***p < 0.01.

5-day, 7-day… 30-day predictions in different time periods. In addition, the values of MAPE, MAE, and its relative improvements are shown in Tables 13–20. All training and testing datasets indicate that the social media data MAPE is smaller than the operational data MAPE; in addition, the social media data MAE is smaller than the operational data MAE. This result exhibits the advantage of our model. In addition, regardless of the time of the study, the relative improvements in the results indicate that using the social media data leads to a significant improvement in the predictions. For example, from 2018-09-28 to 2019-01-28, a MAPE enhancement at a range of 10.49 to 17.53 is achieved by adding the social media data into the operational data for warranty claims. Overall, it was indicated that incorporating the social media data improves prediction of warranty claims significantly from a statistical perspective.

5.2.3.2. Effect of time on the perdition models and dataset Size. This section compares the prediction models' performance at different dataset sizes and different prediction model sizes. In Table 21, we documented the training and testing sets from 2018-1-1 in 2 sizes. These datasets were evaluated on predictive models for the 1-to-7-days-ahead predictions and 1-to-30-days-ahead predictions.

In Table 21, we made 1-to-7-days-ahead predictions and 1-to-30-days-ahead predictions in different dataset sizes. 1-to-7-days - ahead prediction models trained with a data set size of 120 days and evaluated with a data set size of 90 days. The results are shown in Table 21 for 1-day, 4-day and 7-day predictions. In addition, 1-to-30-days-ahead prediction models were trained with a data set size of 240 days and evaluated with a data set size of 120 days. The results in the table for 1-day, 4-day and 7-day predictions it has been shown. In addition, the values of MAPE are shown in Table 21.

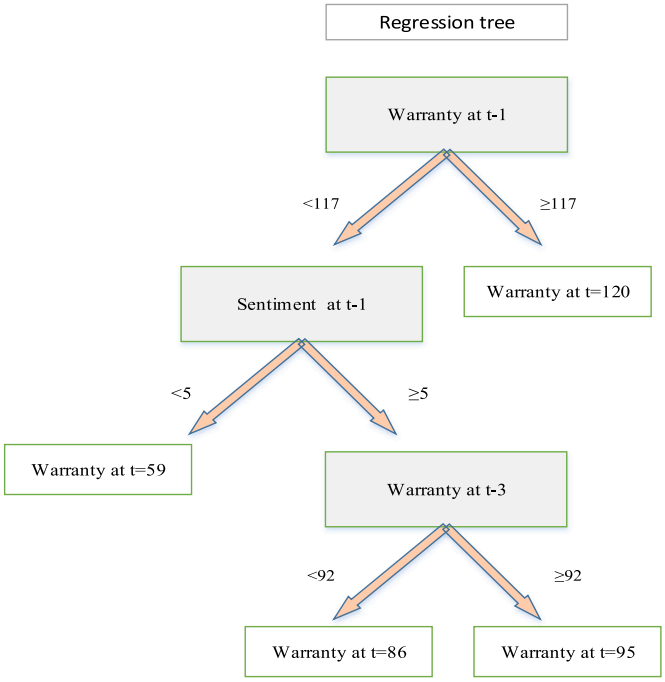


Figure 9. Special example of Regression tree

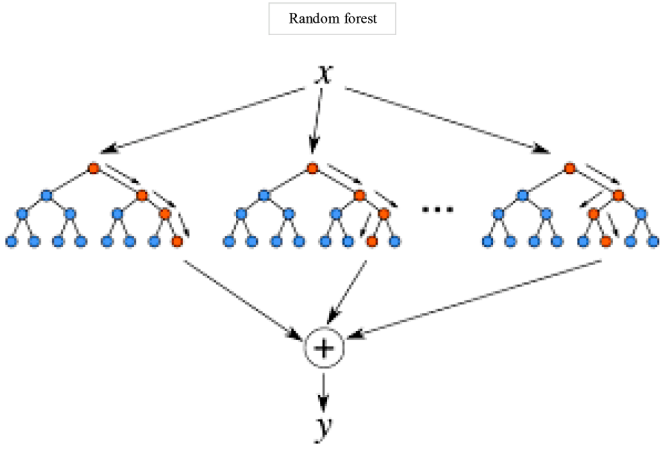


Figure 10. A special example of Random Forecast

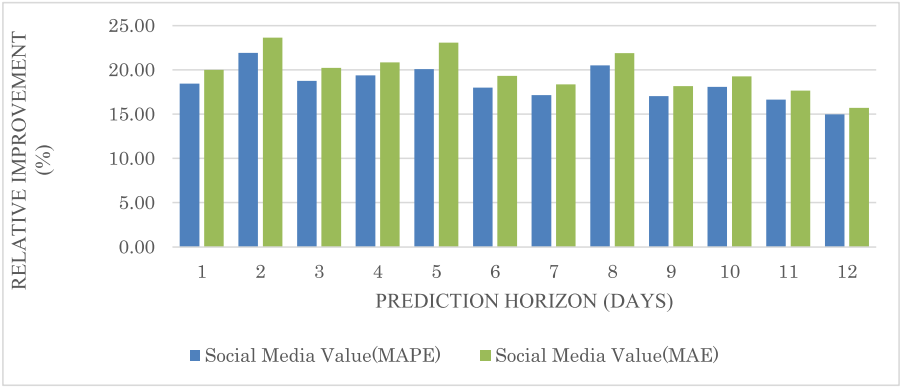


Figure 8. Relative Prediction Improvement over Prediction Horizon

Table 11
Prediction Accuracy Improvement for Returning Customers Prediction Lead Time

Lead Time	Returning Customers MAPE (%)			MAE (Number)		
	Base	Social Media	Relative Improvement (%)	Base	Social Media	Relative Improvement (%)
L=1	15.35	12.94	15.70***	2.32	1.88	19.08
L=2	16.07	13.78	14.25**	2.46	2.03	17.15
L=3	15.51	13.23	14.70**	2.35	1.93	17.82
L=4	16.63	14.21	14.55**	2.56	2.11	17.40
L=5	16.27	14.63	10.08*	2.49	2.19	12.10
L=6	16.91	14.81	12.42**	2.61	2.22	14.80
L=7	17.22	15.05	12.60**	2.67	2.27	14.96
L=8	16.35	14.07	13.94**	2.51	2.09	16.73
L=9	17.94	15.45	13.88***	2.8	2.34	16.36
L=10	18.08	15.84	12.37**	2.83	2.41	14.58
L=15	19.37	17.74	8.42**	3.06	2.76	9.79
L=30	22.81	21.04	7.76	3.7	3.37	8.81

Note: *p <0.1; **p <0.05; ***p < 0.01.

Table 12
Prediction Accuracy Improvement for New Customers Prediction Lead Time

Lead Time	New Customer MAPE (%)			MAE (Number)		
	Base	Social Media	Relative Improvement (%)	Base	Social Media	Relative Improvement (%)
L=1	11.76	9.73	17.26***	6.82	5.52	19.16
L=2	11.25	10.57	6.04*	6.50	6.06	6.74
L=3	11.48	10.96	4.53	6.64	6.31	5.04
L=4	11.88	10.82	8.92**	6.90	6.22	9.89
L=5	12.87	11.08	13.91**	7.54	6.39	15.29
L=6	13.25	11.36	14.26***	7.78	6.57	15.64
L=7	12.59	10.86	13.74**	7.36	6.24	15.14
L=8	13.97	12.49	10.59**	8.25	7.29	11.56
L=9	14.11	12.82	9.14*	8.34	7.51	9.96
L=10	14.38	12.97	9.81**	8.51	7.60	10.67
L=15	14.98	13.88	7.34**	8.90	8.19	7.96
L=30	18.52	16.39	11.50**	11.18	9.81	12.27

Note: *p <0.1; **p <0.05; ***p < 0.01.

6. Research implications

6.1. Management implications

The social media create a robust communication channel between the customers and the manufacturer. The customers can conveniently express their opinions about the products on social media. Consumers post useful suggestions about product design and discuss it with others with the capabilities such as quote, share, and like in social media. In addition, they can explain any problems with product usage or discuss product-related topics with other consumers. In order to perform a more rigorous analysis on the product's usage, managers are using consumers' opinions in social media. In addition, they can categorize and classify consumer's problems with regard to product usage and quality.

It was indicated in the present work that social media data can create a rich database that can be used to predict more accurate results with regard to the number of warranty claims. Predicting warranty claims empowers managers to maximize the return on investment, profits, and efficiency while minimizing inventory level, cost, and customer dissatisfaction. As a result, we highly recommend companies to have a separate social media warranty claim database along with their operational warranty database.

Using low-quality data for the prediction of the number of warranty claims is not very precise because "product's age is not specified", "the

Table 13
Warranty Prediction Improvement of 1-day-head with Various Starting Dates of the Training Set

Predict lead time		L=1 MAPE (%)			MAE (Number)			
Training End	Testing End	Base	Social Media	Relative improvement(%)	Base	Social Media		Relative improvement(%)
8-28	12-29	12.74	10.39	18.45***	12.05	9.64	20.00	
9-28	1-28	12.89	10.96	14.97**	12.20	10.23	16.15	
10-26	2-28	12.61	10.27	18.56***	11.92	9.52	20.13	
11-26	3-30	12.57	10.66	15.19**	11.88	9.92	16.50	
12-26	4-30	12.86	10.98	14.62**	12.17	10.25	15.82	

Note: *p <0.1; **p <0.05; ***p < 0.01.

Table 14
Warranty Prediction Improvement of 3-day-ahead with Various Starting Dates of the Training Set

Predict lead time		L=3 MAPE (%)			MAE (Number)		
Training End	Testing End	Base	Social Media	Relative improvement(%)	Base	Social Media	Relative improvement(%)
8-28	12-29	13.43	10.91	18.76**	12.76	10.18	20.22
9-28	1-28	13.58	10.98	19.15**	12.91	10.25	20.60
10-26	2-28	13.32	10.51	21.10***	12.65	9.77	22.77
11-26	3-30	13.12	10.42	20.58***	12.44	9.67	22.27
12-26	4-30	14.29	11.37	20.43**	13.64	10.65	21.93

Note: *p <0.1; **p <0.05; ***p < 0.01.

Table 15
Warranty Prediction Improvement of 5-day-ahead with Various Starting Dates of the Training Set

Predict lead time		L=5					
		MAPE (%)			MAE (Number)		
Training End	Testing End	Base	Social Media	Relative improvement (%)	Base	Social Media	Relative improvement (%)
8-28	12-29	13.59	10.86	20.09***	12.92	9.94	23.07
9-28	1-28	12.94	10.49	18.93**	12.26	9.75	20.47
10-26	2-28	12.86	10.14	21.15***	12.17	9.39	22.84
11-26	3-30	13.16	10.73	18.47**	12.48	9.99	19.95
12-26	4-30	13.87	11.12	20.55***	13.21	10.39	21.33

Note: *p <0.1; **p <0.05; ***p < 0.01.

Table 16
Warranty Prediction Improvement of 7-day-ahead with Various Starting Dates of the Training Set

Predict lead time		L=7					
		MAPE (%)			MAE (Number)		
Training End	Testing End	Base	Social Media	Relative improvement(%)	Base	Social Media	Relative improvement(%)
8-28	12-29	14.59	12.09	17.14**	13.95	11.39	18.35
9-28	1-28	14.3	11.95	16.43**	13.65	11.24	17.66
10-26	2-28	14.52	12.04	17.08**	13.87	11.33	18.31
11-26	3-30	14.96	12.29	17.85**	14.32	11.59	19.06
12-26	4-30	15.21	12.17	19.99***	14.58	11.47	21.36

Note: *p <0.1; **p <0.05; ***p < 0.01.

Table 17
Warranty Prediction Improvement of 9-day-ahead with Various Starting Dates of the Training Set

Predict lead time		L=9					
		MAPE (%)			MAE (Number)		
Training End	Testing End	Base	Social Media	Relative improvement(%)	Base	Social Media	Relative improvement(%)
8-28	12-29	15.51	12.87	17.02**	14.89	12.18	18.20
9-28	1-28	15.77	13.1	16.93**	15.16	12.33	18.67
10-26	2-28	16.04	13.69	14.65**	15.43	13.02	15.62
11-26	3-30	15.96	13.02	18.42***	15.35	12.34	19.61
12-26	4-30	16.33	14.43	11.64*	15.73	13.78	12.37

Note: *p <0.1; **p <0.05; ***p < 0.01.

Table 18
Warranty Prediction Improvement of 11-day-ahead with Various Starting Dates of the Training Set

Predict lead time		L=11					
		MAPE (%)			MAE (Number)		
Training End	Testing End	Base	Social Media	Relative improvement (%)	Base	Social Media	Relative improvement (%)
8-28	12-29	16.34	13.48	17.15**	15.74	12.81	18.61
9-28	1-28	15.25	13.11	14.03**	14.62	12.43	14.98
10-26	2-28	15.09	12.94	14.25**	14.46	12.26	15.21
11-26	3-30	15.68	13.53	13.71**	15.06	12.86	14.61
12-26	4-30	17.8	13.17	26.01**	17.23	12.49	27.52

Note: *p <0.1; **p <0.05; ***p < 0.01.

Table 19
Warranty Prediction Improvement of 15-day-ahead with Various Starting Dates of the Training Set

predict lead time		L=15					
		MAPE (%)			MAE (Number)		
Training End	Testing End	Base	Social Media	Relative improvement(%)	Base	Social Media	Relative improvement(%)
8-28	12-29	17.02	14.19	16.63**	16.44	13.54	17.64
9-28	1-28	16.96	14.05	17.16**	16.37	13.39	18.20
10-26	2-28	17.2	14.1	18.02**	16.62	13.35	19.68
11-26	3-30	17.59	14.36	18.36***	17.02	13.71	19.45
12-26	4-30	17.37	14.28	17.79***	16.79	13.63	18.85

Note: *p <0.1; **p <0.05; ***p < 0.01.

delays in analyzing data are not determined", and "the usage behavior of the consumer that may have caused a problem is not mentioned." In the presented method, to address these flaws, we suggested using consumers' opinions on social media along with warranty operational data, which leads to more accurate predictions. Furthermore, it is more cost-effective to use social media data than operational data, which is collected by human resources. Social media always enable managers to have real-time information, which leads to fast and accurate decisions

Table 20
Warranty Prediction Improvement of 30-day-ahead with Various Starting Dates of the Training Set

predict lead time		L=30 MAPE (%)			MAE (Number)		
Training End	Testing End	Base	Social Media	Relative improvement(%)	Base	Social Media	Relative improvement(%)
8-28	12-29	20.83	17.71	14.98**	20.34	17.14	15.73
9-28	1-28	20.61	17.53	14.94**	20.11	16.96	15.66
10-26	2-28	20.14	17.19	14.65**	19.63	16.61	15.38
11-26	3-30	21.1	18.02	14.60**	20.61	17.46	15.28
12-26	4-30	20.87	18.57	11.02*	20.38	18.02	11.56

Note: *p <0.1; **p <0.05; ***p < 0.01.

Table 21
Comparing MAPE for the Different Models and Different Dataset Size

MAPE	Testing sets (days)	Training sets (days)	Models L=7		L=4		L=1	
			Social media	Base	Social media	Base	Social media	Base
14.11	16.41	13.79	16.22	12.63	14.25	90	120	1-to-7-days-ahead predictions
12.09	14.59	11.23	13.93	10.39	12.74	120	240	1-to-30-days-ahead predictions

based on the consumers' taste.

The method presented in this paper has a number of limitations. Firstly, the data collected from social media is dirty data, suffering from problems such as spelling or punctuation errors, incorrect data associated with a warranty claim, incomplete data, or even data that have been duplicated on the company's official page. Therefore, in order to analyze this type of data, managers first need to clean the data. For example, when a Twitter user posts a comment about a problem with any products on the company's official page, they browse the company's page to find whether the warranty covers the problem or not; hence, this type of data is associated with a warranty claim. It should be noted that incorrect data associated with a warranty claim can be cleaned through a process known as data cleaning. The second limitation involves the high dimensions of the data, which is needed to be analyzed. However, it is not possible to use plain linear models, including moving-average and autoregressive models, for this type of data. Such problems can be solved using advanced methods to select and classify variables, including machine learning algorithms.

The major managerial contributions of the current study are as follows:

(1) This paper suggested using consumer opinions on social media along with warranty operational data, which can lead to more accurate predictions.
(2) This paper enhances managerial understanding of social media data in the warranty claim services management context and helps managers have a concrete decision-making process in relation to their supply chain management.
(3) This paper helps the management process of warranty services to make financial strategy plans more consumer-focused.
(4) Based on the presented framework, managers can prepare warranty strategy plans to minimize inventory levels and waste while maximizing consumers' satisfaction and profits.

The presented model is based on a daily timeframe, so managers should note that this method cannot be used for long-term predictions, such as six months or more forwards. Besides, we have considered the data for all Samsung smartphones with all the problems that are covered by the warranty; however, companies can use the data for a specific product and a specific policy.

6.2. Theoretical implications

Consumers use their product warranty for many reasons, including breakdown, product unsatisfactory, support, etc. Most consumers find it easier to comment on products with anonymous communication channels [88]. Today, it is widely used by consumers due to its anonymity in social networks. Consumers express their opinions about product warranty on social networks, which can affect the company's reputation. While satisfaction with warranty service can boost a manufacturer's reputation, dissatisfaction with warranty service can lead to a drop in company revenue and brand credibility. Different companies have different strategies to increase customer satisfaction with warranty service in the reverse supply chain. This will not happen unless the companies make accurate predictions about the number of warranty claims. Companies have predicted the number of warranties, and different prediction methods have been developed. However, the most important issue in these methods is the use of operational data. Operational data usually have a low quality [4, 9]. In the presented method, to address these flaws, we suggested using consumers' opinions on social media along with the operational warranty data, which leads to a more accurate prediction. The accuracy in predicting warranty claims in the reverse supply chain increases consumer satisfaction, reduces the costs of the product development process, adjusts financial manufacturer's plans, and so on.

7. Conclusions

This study has examined the value of employing social media data to enhance warranty claim predictions, and it was indicated how integration of social media data into warranty claim prediction leads to considerable developments. The obtained results can be applied to out-of-sample prediction tests for both the full dataset and various data subsamples, including new and returning customers and a variety of training periods. Improving the accuracy of warranty claim prediction brings about fundamental operational advantages in different conditions. It is possible to translate a more accurate prediction into maximized return on investment, profit, and efficiency while minimizing inventory level, cost, customer dissatisfaction, and adjusting ad spent, and so on. Accurate predictions can be extremely valuable for warranty service providers, such as Samsung, Apple, Acer, and so on. It is worth to have better predictions even in the case that the companies cannot affect their supply across the prediction horizon because they can typically address the demand side of their business.

Conversations with the partner of the study highlighted the significant practical insights reflected in the obtained results. As an instance, in the company, longer-term predictions can help the finance departments prepare their financial plans. Our work primarily addresses medium-term predictions; however, a similar methodology can be adopted and

used for longer-term predictions. The present work can be generalized in different ways. Being an aftersales service provider, they continuously improve their communication efforts on Twitter as well as other social media platforms. The improvements attained by the incorporation of social media data into the processes of prediction were both surprising and interesting for them. When the obtained results were shared, some processes were considered for systematic tracking of the social media data aimed at feeding their prediction model.

Addressing the data of social media can be challenging. One of the significant challenges faced by practitioners and scholars is the unstructured, textual nature of a considerable part of social media data, showing differences with the sort of data addressed by the academics and practitioners conventionally. As an instance, natural language processing methods were employed in the present work for encoding the sentiments of the comments that users made on social media. After overcoming the primary challenges, such data can be potentially employed in new ways, leading to more efficiency of the firms, and delivering the customers better experiences. More time is required until other firms also take measures to employ such data for routine improvement of their assortment decisions and enhancement of their customers' experiences.

This paper has only considered the problem of one-to-thirty-days-ahead predictions for warranty claims. Based on the warranty suppliers/manufacturers' viewpoint, long-term predictions (e.g., 6, 24, and 36 months, over even longer) of warranty claims are of critical importance in their financial planning. Long-term predictions for warranty claims will be covered in future studies using the available social media data. The Data collected from social media are comments under Samsung posts on its official Twitter page. These comments generally fall into two categories. These categories are (1) customers start commenting and receiving help from other Twitter users or the official account before using the warranty claim service (2) customers start commenting after using the warranty claim service due to their satisfaction or dissatisfaction with the quality of service they received. In the second category, we have a delay in receiving feedback from customers. This is another limitation of our research, which could be considered in future research. Since we are having a high-dimensional problem, we can choose different ways in machine learning to overcome this challenge. For example, Feature selection methods, Regularized Algorithms, Projection Methods, and so on. In this paper, we have used the Random Forrest algorithm and shown the results. Other methods, such as Elastic Net Algorithm, which is a Regularization algorithm, perform very well on this type of data. Therefore, this algorithm can be examined in future research. It is hoped that the present project can inspire other researchers to carry out more research with the use of structured as well as unstructured data provided by social media platforms.

Author statements

1- Sajjad Shokouhyar
2- Sadra Ahmadi
3-mahdi ashrafzadeh
Conceptualization: 1
Data curation: 1,2,3
Formal analysis: 1,3
Funding acquisition: 2
Investigation: 2,3
Methodology: 1,3
Project administration: 2
Resources: 1,2
Software: 3
Supervision: 2
Validation: 1,2
Visualization: 3
Roles/Writing - original draft: 1,2
Writing - review & editing: 1,2,3

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Table 6, Table 8.

References

[1] Xie W, Shen L, Zhong Y. Two-dimensional aggregate warranty demand forecasting under sales uncertainty. IISE Transactions 2017. https://doi.org/10.1080/24725854.2016.1263769.
[2] Karim MR, Suzuki K. Analysis of warranty claim data: A literature review. In: International Journal of Quality and Reliability Management; 2005. https://doi.org/10.1108/02656710510610820.
[3] Kleyner A, Sandborn P. Forecasting the cost of unreliability for products with two-dimensional warranties. In: Proceedings of the European Safety and Reliability Conference 2006, ESREL 2006 - Safety and Reliability for Managing Risk; 2006.
[4] Wu S. Warranty data analysis: A review. In: Quality and Reliability Engineering International; 2012. https://doi.org/10.1002/qre.1282.
[5] Wu S, Akbarov A. Support vector regression for warranty claim forecasting. European Journal of Operational Research 2011. https://doi.org/10.1016/j.ejor.2011.03.009.
[6] Wu S, Akbarov A. Forecasting warranty claims for recently launched products. Reliability Engineering and System Safety 2012. https://doi.org/10.1016/j.ress.2012.06.008.
[7] Ye ZS, Murthy DNP. Warranty menu design for a two-dimensional warranty. Reliability Engineering and System Safety 2016. https://doi.org/10.1016/j.ress.2016.05.013.
[8] Dai A, Zhang Z, Hou P, Yue J, He S, He Z. Warranty Claims Forecasting for New Products Sold with a Two-Dimensional Warranty. Journal of Systems Science and Systems Engineering 2019. https://doi.org/10.1007/s11518-019-5434-8.
[9] Wu S. A review on coarse warranty data and analysis. In: Reliability Engineering and System Safety; 2013. https://doi.org/10.1016/j.ress.2012.12.021.
[10] Bian Y, Yan S, Zhang W, Xu H. Warranty strategy in a supply chain when two retailer's extended warranties bundled with the products. Journal of Systems Science and Systems Engineering 2015. https://doi.org/10.1007/s11518-015-5270-4.
[11] Dhir A, Buragga K, Boreqqah AA. Tweeters on campus: Twitter a learning tool in classroom? Journal of Universal Computer Science 2013.
[12] He W, Zhang W, Tian X, Tao R, Akula V. Identifying customer knowledge on social media through data analytics. Journal of Enterprise Information Management 2019. https://doi.org/10.1108/JEIM-02-2018-0031.
[13] Othman R, Belkaroui R, Faiz R. Extracting Product Features for Opinion Mining Using Public Conversations in Twitter. Procedia Computer Science 2017. https://doi.org/10.1016/j.procs.2017.08.122.
[14] Zengin Alp Z, Gündüz Öğüdücü Ş. Identifying topical influencers on twitter based on user behavior and network topology. Knowledge-Based Systems 2018. https://doi.org/10.1016/j.knosys.2017.11.021.
[15] Cui R, Gallino S, Moreno A, Zhang DJ. The Operational Value of Social Media Information. Production and Operations Management 2018. https://doi.org/10.1111/poms.12707.
[16] Yang D, He Z, He S. Warranty claims forecasting based on a general imperfect repair model considering usage rate. Reliability Engineering and System Safety 2016. https://doi.org/10.1016/j.ress.2015.09.012.
[17] Heydari J, Choi TM, Radkhah S. Pareto improving supply chain coordination under a money-back guarantee service program. Service Science 2017. https://doi.org/10.1287/serv.2016.0153.
[18] Lee, K., & Khan, M. A. (2012). Exploring the Impacts of Service Guarantee Strategy. https://doi.org/10.1080/10548408.2012.648530, 29(2), 133–146. https://doi.org/10.1080/10548408.2012.648530.
[19] Wang X, Liu B, Zhao X. A performance-based warranty for products subject to competing hard and soft failures. International Journal of Production Economics 2021;233. https://doi.org/10.1016/j.ijpe.2020.107974.
[20] Alqahtani AY, Gupta SM. Money-back guarantee warranty policy with preventive maintenance strategy for sensor-embedded remanufactured products. Journal of Industrial Engineering International 2018. https://doi.org/10.1007/s40092-018-0259-5.
[21] Huang Z, Feng T. Money-back guarantee and pricing decision with retailer's store brand. Journal of Retailing and Consumer Services 2020;52(December 2018): 101897. https://doi.org/10.1016/j.jretconser.2019.101897.
[22] Heiman A, McWilliams B, Zilberman D. Demonstrations and money-back guarantees: Market mechanisms to reduce uncertainty. Journal of Business Research 2001. https://doi.org/10.1016/S0148-2963(00)00181-8.
[23] Moorthy S, Srinivasan K. Signaling Quality with a Money-Back Guarantee: The Role of Transaction Costs. Marketing Science 1995. https://doi.org/10.1287/mksc.14.4.442.

[24] Chiarello F, Bonaccorsi A, Fantoni G. Technical Sentiment Analysis. Measuring Advantages and Drawbacks of New Products Using Social Media. Computers in Industry 2020. https://doi.org/10.1016/j.compind.2020.103299.

[25] Li X, Law R, Xie G, Wang S. Review of tourism forecasting research with internet data. Tourism Management 2021;83(October 2020):104245. https://doi.org/10.1016/j.tourman.2020.104245.

[26] Boone T, Ganeshan R, Hicks RL, Sanders NR. Can Google Trends Improve Your Sales Forecast? Production and Operations Management 2018. https://doi.org/10.1111/poms.12839.

[27] Boone T, Ganeshan R, Jain A, Sanders NR. Forecasting sales in the supply chain: Consumer analytics in the big data era. International Journal of Forecasting 2019. https://doi.org/10.1016/j.ijforecast.2018.09.003.

[28] Dawson P, Carless D, Lee PPW. Authentic feedback: supporting learners to engage in disciplinary feedback practices. Assessment and Evaluation in Higher Education 2021;46(2):286–96. https://doi.org/10.1080/02602938.2020.1769022.

[29] Surucu-Balci E, Balci G, Yuen KF. Social Media Engagement of Stakeholders: A Decision Tree Approach in Container Shipping. Computers in Industry 2020. https://doi.org/10.1016/j.compind.2019.103152.

[30] Wang L, Lee JH. The impact of K-beauty social media influencers, sponsorship, and product exposure on consumer acceptance of new products. Fashion and Textiles 2021;8(1). https://doi.org/10.1186/s40691-020-00239-0.

[31] Choi TM, Guo S, Luo S. When blockchain meets social-media: Will the result benefit social media analytics for supply chain operations management? Transportation Research Part E: Logistics and Transportation Review 2020. https://doi.org/10.1016/j.tre.2020.101860.

[32] Devi Y, Ganguly KK. Social media in operations and supply chain management: A systematic literature review to explore the future. Operations and Supply Chain Management 2021;14(2):232–48. https://doi.org/10.31387/oscm0450299.

[33] Orji IJ, Kusi-Sarpong S, Gupta H. The critical success factors of using social media for supply chain social sustainability in the freight logistics industry. International Journal of Production Research 2020. https://doi.org/10.1080/00207543.2019.1660829.

[34] Li Shanshan, He Y. Compensation and information disclosure strategies of a green supply chain under production disruption. Journal of Cleaner Production 2021;281:124851. https://doi.org/10.1016/j.jclepro.2020.124851.

[35] Miao S, Wang T, Chen D. System dynamics research of remanufacturing closed-loop supply chain dominated by the third party. Waste Management and Research 2017. https://doi.org/10.1177/0734242X16684384.

[36] Nasrollahi M. The impact of Firm's Social Media Applications on Green Supply Chain Management. International Journal of Supply Chain Management 2018.

[37] Sujata M, Khor KS, Ramayah T, Teoh AP. The role of social media on recycling behaviour. Sustainable Production and Consumption 2019. https://doi.org/10.1016/j.spc.2019.08.005.

[38] Yunanto R. Android-based Social Media System of Household Waste Recycling: Designing and User Acceptance Testing. In: IOP Conference Series: Materials Science and Engineering; 2018. https://doi.org/10.1088/1757-899X/407/1/012139.

[39] Mircică N. Restoring public trust in digital platform operations: Machine learning algorithmic structuring of social media content. Review of Contemporary Philosophy 2020. https://doi.org/10.22381/RCP1920209.

[40] Sharifi Z, Shokouhyar S. Promoting consumer's attitude toward refurbished mobile phones: A social media analytics approach. Resources, Conservation and Recycling 2021;167(March):105398. https://doi.org/10.1016/j.resconrec.2021.105398.

[41] Zheng L, He Z, He S. A novel probabilistic graphic model to detect product defects from social media data. Decision Support Systems 2020. https://doi.org/10.1016/j.dss.2020.113369.

[42] Peng T, Chunling L. Designing differential service strategy for two-dimensional warranty based on warranty claim data under consumer-side modularisation. In: Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability; 2020. https://doi.org/10.1177/1748006X19886162.

[43] Alkahtani M, Choudhary A, De A, Harding JA. A decision support system based on ontology and data mining to improve design using warranty data. Computers and Industrial Engineering 2019. https://doi.org/10.1016/j.cie.2018.04.033.

[44] Stevens RE. The Marketing Research Guide. Routledge; 2006.

[45] Wang X, Xie W. Two-dimensional warranty: A literature review. In: Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability; 2018. https://doi.org/10.1177/1748006X17742776.

[46] Nagahisarchoghaei M, Dodd J, Nagahi M, Ghanbari G, Poudyal S. Analysis of a Warranty-Based Quality Management System in the Construction Industry. In: *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020, January*; 2020. https://doi.org/10.1109/ICDABI51230.2020.9325692.

[47] Jindal A, Sangwan KS. Evaluation of collection methods in reverse logistics by using fuzzy mathematics. Benchmarking: An International Journal 2015;22(3):393–410. https://doi.org/10.1108/BIJ-05-2013-0062.

[48] Li Suhong, Ragu-Nathan B, Ragu-Nathan TS, Subba Rao S. The impact of supply chain management practices on competitive advantage and organizational performance. Omega 2006;34(2):107–24. https://doi.org/10.1016/J.OMEGA.2004.08.002.

[49] Lancioni RA, Smith MF, Oliva TA. The role of the internet in supply chain management. Industrial Marketing Management 2000. https://doi.org/10.1016/s0019-8501(99)00111-x.

[50] Murthy DNP, Djamaludin I. New product warranty: A literature review. International Journal of Production Economics 2002. https://doi.org/10.1016/S0925-5273(02)00153-6.

[51] Zhang Y, He Z, He S, Cai K, Wang D. Manufacturer warranty service outsourcing strategies in a dual-channel supply chain. International Transactions in Operational Research 2020;27(6):2899–926. https://doi.org/10.1111/ITOR.12769.

[52] Rai BK. Warranty spend forecasting for subsystem failures influenced by calendar month seasonality. IEEE Transactions on Reliability 2009. https://doi.org/10.1109/TR.2009.2019673.

[53] Wifvat K, Kumerow J, Shemyakin A. Copula model selection for vehicle component failures based on warranty claims. Risks 2020. https://doi.org/10.3390/risks8020056.

[54] Rai B, Singh N. Forecasting warranty performance in the presence of the "maturing data" phenomenon. International Journal of Systems Science 2005. https://doi.org/10.1080/00207720500139930.

[55] Hrycej T, Grabert M. Warranty cost forecast based on car failure data. In: IEEE International Conference on Neural Networks - Conference Proceedings; 2007. https://doi.org/10.1109/IJCNN.2007.4370939.

[56] Stamenkovic DD, Popovic VM. Warranty optimisation based on the prediction of costs to the manufacturer using neural network model and Monte Carlo simulation. International Journal of Systems Science 2015. https://doi.org/10.1080/00207721.2013.792972.

[57] Chen J, Chen T. Research on classification method of spare parts inventory based on warranty data. In: Proceedings - 2016 IEEE International Conference on Service Operations and Logistics, and Informatics, SOLI 2016; 2016. https://doi.org/10.1109/SOLI.2016.7551686.

[58] Darmawan MF, Jamahir NI, Saedudin RR, Kasim S. Comparison between ANN and multiple linear regression models for prediction of warranty cost. International Journal of Integrated Engineering 2018. https://doi.org/10.30880/ijie.2018.10.06.027.

[59] Ferencek A, Borštnar MK, Kofjač D, Škraba A, Sašek B. Deep learning predictive models for terminal call rate prediction during the warranty period. Proceedings of the 15th International Symposium on Operations Research, SOR 2019 2019;11(2):11–6. https://doi.org/10.2478/bsrj-2020-0014.

[60] Khoshkangini R, Pashami S, Nowaczyk S. Warranty claim rate prediction using logged vehicle data. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2019. https://doi.org/10.1007/978-3-030-30241-2_55.

[61] Torgunov D, Trundle P, Campean F, Neagu D, Sherratt A. Vehicle Warranty Claim Prediction from Diagnostic Data Using Classification. Advances in Intelligent Systems and Computing 2020. https://doi.org/10.1007/978-3-030-29933-0_40.

[62] Khoshkangini R, Mashhadi PS, Berck P. Early Prediction of Quality Issues in Automotive Modern Industry 2020:1–15. https://doi.org/10.3390/info11070354.

[63] Hsu CC, Lee YC, Lu PE, Lu SS, Lai HT, Huang CC, Wang C, Lin YJ, Su WT. Social media prediction based on residual learning and random forest. In: MM 2017 - Proceedings of the 2017 ACM Multimedia Conference; 2017. https://doi.org/10.1145/3123266.3127894.

[64] Chen D, Zhang D, Tao F, Liu A. Analysis of Customer Reviews for Product Service System Design based on Cloud Computing. Procedia CIRP 2019;83:522–7. https://doi.org/10.1016/J.PROCIR.2019.03.116.

[65] Fabijan A, Olsson HH, Bosch J. Customer Feedback and Data Collection Techniques in Software R&D: A Literature Review. Lecture Notes in Business Information Processing 2015;210:139–53. https://doi.org/10.1007/978-3-319-19593-3_12.

[66] Fariborzi E. E-mail Marketing: Advantages, Disadvantages andImproving Techniques. International Journal of E-Education, e-Business, e-Management and e-Learning 2012;2(3):1–5. https://doi.org/10.7763/ijeeee.2012.v2.116.

[67] Kohavi R. Mining e-commerce data: The good, the bad, and the ugly. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2001. p. 8–13.

[68] D. BS, Gore D. Research Article Open Access Sentiment Analysis On Twitter Data Using Support Vector Machine. International Journal of Computer Science Trends and Technology (IJCST) 2016;4(3):831–7.

[69] Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. In: EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference; 2013.

[70] Amirmokhtar Radi S, Shokouhyar S. Toward consumer perception of cellphones sustainability: A social media analytics. Sustainable Production and Consumption 2021;25:217–33. https://doi.org/10.1016/j.spc.2020.08.012.

[71] Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: International Joint Conference of Artificial Intelligence; 1995.

[72] Osadchiy N, Gaur V, Seshadri S. Sales forecasting with financial indicators and experts' input. Production and Operations Management 2013. https://doi.org/10.1111/poms.12022.

[73] Casella, G., Fienberg, S., & Olkin, I. (2013). *An Introduction to Statistical Learning with Applications in R* (Vol. 102). https://doi.org/10.1016/j.peva.2007.06.006.

[74] Kuhn M, Johnson K. Applied predictive modeling. Applied Predictive Modeling 2013:1–600. https://doi.org/10.1007/978-1-4614-6849-3.

[75] Azadi S, Karimi-jashni A. Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate : A case study of Fars province, Iran. Waste Management 2015. https://doi.org/10.1016/j.wasman.2015.09.034.

[76] Ji Z, Jiao F, Pang Y, Shao L. Deep Attentive and Semantic Preserving Video Summarization. Neurocomputing 2020. https://doi.org/10.1016/j.neucom.2020.04.132.

[77] Sreeraj M, Joy J, Jose M, Varghese M, Rejoice TJ. Comparative analysis of Machine Learning approaches for early stage Cervical Spondylosis detection. Journal of King Saud University - Computer and Information Sciences 2020. https://doi.org/10.1016/j.jksuci.2020.08.010.

[78] de Myttenaere A, Golden B, Le Grand B, Rossi F. Mean Absolute Percentage Error for regression models. Neurocomputing 2016;192:38–48. https://doi.org/10.1016/j.neucom.2015.12.114.
[79] Jiu-xun, S., & Walker, S. (2017). *Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error*.
[80] M Bland. (2015). *An Introduction to Medical Statistics - Martin Bland - Google Books*. https://books.google.com/books?hl=en&lr=&id=fKgXCgAAQBAJ&oi=fnd&pg=PP1&dq=An+introduction+to+medical+statistics.&ots=Ex2VlIISkq&sig=jUozNNX_-Jo_ZkmLWQFkEnSBtqE#v=onepage&q=An introduction to medical statistics.&f=false.
[81] Runge, J., & Zmeureanu, R. (2021). *A Review of Deep Learning Techniques for Forecasting Energy Use in Buildings*.
[82] Lei C, Deng J, Cao K, Ma L, Xiao Y, Ren L. A random forest approach for predicting coal spontaneous combustion. Fuel 2018;223(December):63–73. https://doi.org/10.1016/j.fuel.2018.03.005.
[83] Li K, Niskanen J, Kolehmainen M, Niskanen M. PT US CR. Expert Systems With Applications 2016. https://doi.org/10.1016/j.eswa.2016.05.029.
[84] Munkhdalai, L., Ryu, K. H., & Namsrai, O. (2021). *applied sciences A Partially Interpretable Adaptive Softmax Regression for Credit Scoring*.
[85] Huang, F., Chen, J., Lin, Z., Kang, P., & Yang, Z. (2018). Random forest exploiting post-related and user-related features for social media popularity prediction. MM 2018 - Proceedings of the 2018 ACM Multimedia Conference. https://doi.org/10.1145/3240508.3266439.
[86] Clement, J., 2019. Number of Monthly Active Twitter Users Worldwide from 1st quarter 2010 to 1st quarter 2019 (in millions).
[87] Statista. 2020. Statista - The Statistics Portal. [online] Available at: [Accessed 19 June 2020].
[88] Denzil, C., A. S. Leandro, M. Mainack, B. Fabrício, and P. G. Krishna. 2015. "The Many Shades of Anonymity: CHARACTERIZING Anonymous Social Media Contents." The 9th International AAAI Conference on Web and Social Media. Oxford, UK. May 26–29, 2015.