AUTOMATED CLASSIFICATION OF WEB CONTENTS IN B2B MARKETING

Software Engineering
Master's Degree Programme in Information and Communication Technology
Department of Computing, Faculty of Technology
Master of Science in Technology Thesis

Author:
Nischal Guragain

Supervisors:
Dr. Bikesh Raj Upreti (Aalto University, School of Business)
Dr. Tapio Pahikkala (University of Turku)

July 2021

Automated classification of web content in B2B marketing
Department of Computing, Faculty of Technology
University of Turku
Subject: Software Engineering
Programme: Master's Degree Programme in Information and Communication Technology
Author: Nischal Guragain
Number of pages: 46 pages
Date: May 2021

Abstract.

Recent growth in digitization has affected how customers seek the information they need to make a purchase decision. This trend of customers making their purchase decision based on the information they collect online is increasing. To accommodate this change in purchase behavior, companies tend to share information about themselves and their products online, which in turn drives the amount of unstructured data produced. To get value from this huge amount of data being produced, the unstructured data needs to be processed before being used in digital marketing applications. When it comes to the companies in the area of business-to-customers (B2C), plenty of research exists on how the digital content can be used for marketing, but for the companies serving businesses (B2B) a huge research gap presides. B2C and B2B marketing might share some analytical concepts but they are different domains. Not much research has been done in the field of machine learning applications within B2B digital marketing. The lack of availability of labeled text data from the B2B domain makes it challenging for researchers to develop text classification models, while several methods have been proposed and used to classify unstructured text data in marketing and other domains.

This thesis builds upon previous works in the field of text classification in general, focusing on the marketing domain, and compares these methods across the dataset available for this research. Text classification methods such as Random Forest, Linear SVM, KNN, Multinomial Naïve Bayes, and Multinomial Logistic Regression dominates the research field, hence these methods are evaluated for the B2B text content classification. In the used dataset, Random Forest Classifier performed best with an average accuracy of 0.85 in the designed five-class classification task.

KEYWORDS: Text Classification, Natural Language Processing, B2B web contents, Digital marketing, B2B marketing

LANGUAGE: English

Table of Contents

1 Introduction

Producing, recording, transferring text data has always been fundamental in preserving information through generations. The process of how the text data is recorded has come a long way over time, from carvings in the stones to documenting in papers and now being made available in digital platforms. Along with the increase in the availability of text data, processing it for information gain has grown as an essential task across different domains. Due to exponential growth in the availability of text data, manually collecting and processing textual data has become next to impossible. Thus, there is a pressing need to classify texts automatically. Due to significant development in computational power and success in different machine learning algorithms and techniques, this challenging task of managing large volumes of digital text data has become achievable now. As a result, companies are investing in the digitalization process driven with various objectives such as gaining customer insights and developing predictive analytics. For example, collecting data from the web and analyzing, which involves collecting web contents and browsing sessions in the form of cookies that includes timestamps and IP addresses helps in building a model to analyze visiting pattern for customers in their websites. Companies then use the collected data and gathered insights to determine the buying behavior of the customer to make a more targeted marketing approach. [43]

In the context of B2B purchase and sales, at present, most of the information gathering is done digitally compared to a more direct approach used earlier. Not that long ago, companies relied on door-to-door marketing and trade fairs

to advertise their products [45]. But now, making product list and details available digitally is fundamental for B2B sales. Survey by Schwartz and Kim (2012) has shown that 70% of buyers begin searching their solutions from google searches [1]. Customers make initial search and product recognition digitally even before making the first contact with the sales companies [1]. It is estimated that roughly 60% of the buyers already make their final choice before making the first contact with the selling companies [1]. Thus, providing the right amount of information at different phases of purchase is vital for successful sales process. For example, for a buyer who is in the initial phase of purchasing a list of products might be enough. The detailed description of the products is viewed when the buyer is in the advanced phase of making decision. The challenge is enormous for the B2B companies as the research on how to utilize the available data to best serve the customers is limited, as most of the research is focused on B2C sales processes.

The technological advancement has transformed buying process and highlighted the importance of digital content in the B2B sales process as well. The need for better digital marketing in the B2B domain is significant to adapt to this change in buying process and behavior. One of the major problems B2B companies face in digital marketing is the lack of research on which methods serve best in using the vast amounts of text data at their disposal. This challenge makes the need for modern Natural Language Processing methods in the B2B sales domain even more vital.

Digital content from B2B websites can be categorized into various classes depending on the motive and target audience. Some classes such as products and careers are more obvious, but it is challenging to accommodate all the contents in some fixed number of classes. Analyzing from marketing

perspective, for example, web pages with the list of products, detailed product pages, and pages that provide information of the company to the general public and stakeholders are more significant than the one where the vacancy are posted. While it is clear that those different companies and digital content creators emphasize different dimensions of B2B web contents, yet there is a lack of research in understanding the nature of the text contents. This thesis first categorizes B2B web contents into different classes, based on the interview with a digital content creator from one of the major B2B companies in Europe, develops an automated model to classify the contents, and then analyzes subsequent prediction results.

1.1 Research Statement and Thesis structure

In B2B sales gathering knowledge on customers' interests is one of the most important aspects. Customers browse through different web pages before making a final purchase decision. They embark on a sort of a journey from product recognition, to detailed information of product, and ultimately to the final purchase. In this journey, the customers learn as much as they can about the products and also about the seller companies. Precisely, the focus of the B2B sellers is on the tools pushing the right information to the potential clients to facilitate their buying journey. To do that, it is essential for the companies to understand types of contents the customers are browsing at a particular phase of their buying journey. [3]

In the Finnish marketing environment, the art of creating digital content has taken a huge stride to match with the changing behaviors and preferences of B2B customers. While the digital content creators know their audience and their target message, there is still some gaps that remains to be filled i.e. the structured way of publishing and classifying the created digital contents [4].

Lack of labeled data in the B2B domain makes it difficult for the marketers to understand the digital contents' categories and in turn, they are left to guess while trying to identify the customers' information need at various decision-making stages. There lies a significant research gap to fill this void and there is a need to develop a framework for classifying available B2B digital text contents.

To fulfill the main objective of this thesis, previous works in the field of multiclass classifications are analyzed, with a particular interest in the business domain. The review of previous work shows that despite the information overload in the B2B domain, application of text mining in order to utilize the ever-growing data are still scarce. This opens for a wide range of possibilities for future research. While the scope is great, this particular research is limited to:

- Formulating the classification rules for the classifying digital B2B web content.
- Creating training data based on the classification rules.
- Framework for classifying B2B website contents.
- Validating the classification rule by using the intercoder reliability testing method.
- Training the classification models.
- Discussing the results.

2 Related Work

Fabrizio Sebastiani was the pioneer in mechanizing automated classification of texts into predefined classes, and he explains that due to the overwhelming growth of digital accessibility in the research paper, Text categorization in Encyclopedia of Database Technologies and Applications (2005). This approach has seen wide adoption and furthermore, he explains that the volume of digital documents generated will grow exponentially and machine learning techniques are the predominant way to handle these documents. Sebastiani explains the need for a process that naturally builds a classification model by learning from a set of manually labeled text documents with an eye for future use. [5]

Text classification has seen a major stride from being a minor research field in the late '80s to a completely innovated investigation field, which produces insightful, powerful, and generally valuable results with many application areas. This achievement has been powered mainly by two factors: first, the growing interests of researchers in textual data, which has lately brought about the utilization of the latest machine learning approaches in text classification applications, and second, the accessibility of standard benchmarks, (for example, Reuters-21578 and OHSUMED), that has supported researchers on setting a target that could be contrasted with each other, and in which the best techniques and calculations could emerge. [6]

As the creation of text data shifted towards digital medium, researchers also have started focusing on the field of website text classification. At the beginning of digital texts classification works, predominant research were focused on determining whether the websites are phishing websites or not [7] [8]. Different methods have been proposed for the automatic classification of

websites' content as there is a common belief that the manual classification of websites textual content is a huge task due to the explosive rate of the growth of the worldwide web (www). Slowly, the application of websites' textual classification has expanded towards other domains as well, for example, news feeds filtering, opinion filtering, spam filtering, business insights, clinical studies, and recommender systems in e-commerce websites [7]. The growth in application areas also coincides with the growth in the techniques to achieve automation in classifications tasks.

Ferenc Bodon proposed a fast implementation of APRIORI method proposed by Agrawl and Shrikant in the research, A fast APRIORI implementation, which uses multiple scans in the training data sets to discover the frequent itemset. The algorithm scans the transactional database looking for the K-items belonging to the set of items "i". The algorithm uses the previous knowledge to sort the frequent itemset, and a complete scan of the database is necessary for each iteration. In each iteration, the algorithm scans the large itemset discovered in the previous iteration, and discards the infrequent items. This process lasts till no new large itemset is generated. [9] Although APRIORI was initially developed for basket analysis, it relies on statistical features of the data hence produces good results in text classifications [59].

Similarly, in his article, Knowledge-Based Neural Network for Text Classification, 2005, ,Ram Dayal Goyal, described a method that combines Naïve Bayesian text classification and neural network to achieve maximum accuracy while handling the problem of imbalanced and noisy training data. In the proposed method, Goyal constructed ten different training and validation sets using the Reuters dataset. One of the training sets was used to train a Naïve Bayes classifier, and then learned probabilities are the fed to the

neural network. The remaining training sets are used for finding better-generalized accuracy over validation sets via backpropagation. [10]

Tasci and Gungor in their article, LDA-Based keyword selection in text categorization, described the importance of reducing the dimensionality in text classification [11]. They compared more traditional feature selection methods to a generative graphical model called Latent Dirichlet Allocation (LDA) that can be used to model and discover underlining structures to the textual data. In the study, Tasci and Gungor used SVM as a classifier and found that information gained performs best on the keywords while LDA-based matric performs similar to the document frequency thresholding. [11]

Over the years, text classification has extended to a wider application areas. In 2009, Yin et al. used supervised learning methods to detect harassment in web-based community platforms [12]. They used text data from "Kongregate", "Myspace", and "Slashdot" to perform the study in which they combined local features, sentiment features, and contextual features to train the model for detecting online harassment. In the study, Yin et al. found that the model that combines TFIDF with sentiment and contextual feature attributes performed better in detecting online harassment than other baseline methods, including the TFIDF method alone. Continuing in the field of online harassment, Nobata et al. 2016 trained an initial model to detect abusive language combining content features, sentiment features and contextual features of documents with years of labeled user comments from yahoo news in an attempt to find a method for detecting hate speech in web-based platforms. Their study found that a combination of word, n-grams, linguistic, syntactic, and distributional semantic features resulted in better accuracy. [12]

Khan et al. (2009) reviewed machine learning algorithms for text-document classifications focusing on text representation and machine learning techniques. In their analysis of feature selection methods, the authors found that Gini-index and Chi-square statistics are the most commonly used and better-performing feature selection methods. There are other feature selection methods used as a single or as hybrid techniques that show promising results and need further explorations. The authors stated that SVM has been one of the most used text classification methods, although it has some difficulties in parameter tuning and kernel selections. [13]

Mikolov et al. introduced an efficient method that can learn word vectors from up to 1.6 billion words of data in a day. The main difference of the proposed skip-gram model and the previously popular neural network architectures was its lack of dense matrix multiplications, which helped in reducing the training time. [14]

Le and Mikolov in 2014 proposed a paragraph vector for performing text classification and sentiment analysis. The paragraph vector learns from vector representation from variable length text pieces such as sentences, paragraphs, and documents. In order to gain better performance in the text classification, the vector representation is learned from the surrounding words in contexts. According to Le and Mikolov, this method helps to overcome one of the main disadvantages of the Bag-Of-Words model in which the order of words is lost during the representation and thus, helps in overcoming the possible constraints of the model. [15]

In 2017, Arusada et al. used customer complaints on Twitter to test the optimal training data creating strategy. They used Naïve Bayes and Support Vector Machine classifiers to test their hypothesis. They described that in

SVM, the key is to determine the optimal boundaries between the classes. They noted that creating optimal training data takes iterations, but a high level of accuracy is possible. [16]

As we can see from above, the application areas of the text classification field had evolved to wider areas and is growing. In the research, the B2B knowledge gap, Wang and Wang noted that most of the research and analytics are focused on the B2C domain [50]. One of the major application areas of machine learning in marketing is within the content marketing. The main idea behind content marketing is to create contents that are considered either valuable or interesting by the potential customers. The content creators then deliver the created content to the potential buyer at the right time to facilitate their buying journey. [52]

In the study involving cases from 22 different companies, Dzyabura et al. used decision trees and SVM to develop prediction models for marketing purpose. The researchers created simple decision trees to predict which consumers will be attracted towards which product based on the previouse purchases. They noted that both models have their advantages. [53]

Cui et al. studied direct marketing responses using large marketing datasets and machine learning models for their research in 2006. They noted that using Bayesian networks had distinct advantages over the other methods in terms of accuracy, speed, and interpretability of the results. They concluded that using machine learning techniques helps the researchers to gain insights from a large and noisy database. [54]

L Ma et al. in their paper related to Machine learning and AI in marketing studied how the development of machine learning techniques has changed the

approach of B2C marketers. They explained in detail the advantages of machine learning algorithms like linear SVM, Naïve Bayes, and artificial neural networks on training and deploying marketing models on various marketing datasets. [55]

3 Text Classification

The process of assigning a predefined label to the text automatically is text classification. In web content classification, text classification is the process of classifying text contents from the web pages into predefined classes. The amount of digital content has increased significantly in the past decade, and manually classifying them has become a next to impossible task. Automating the classification helps in managing large volumes of text documents while keeping them structured. The recent development in computational power means now more texts can be processed in less time and faster compared to processing it manually. [18]

The need to classify text documents was already noted in the early 80s. In order to classify text contents, an "expert system" was created which consisted of a simple if-else rule [19]. Now, the main approaches in natural language processing for text document processing are supervised semi-supervised and unsupervised learning. In supervised learning, the classification model is trained with labeled training data, while for unsupervised learning, training data can be unlabeled. Semi-supervised learning is a technique where a small amount of labeled training data is combined with a large amount of unlabeled data to train the classification model. One of the typical applications of supervised text classification can be found in sentiment analysis, where textual data, mostly user generated textual content, are classified as either positive, negative or neutral. Such classification tasks are often referred to as single label classification where the content belongs to just one of the possible classes. [20]

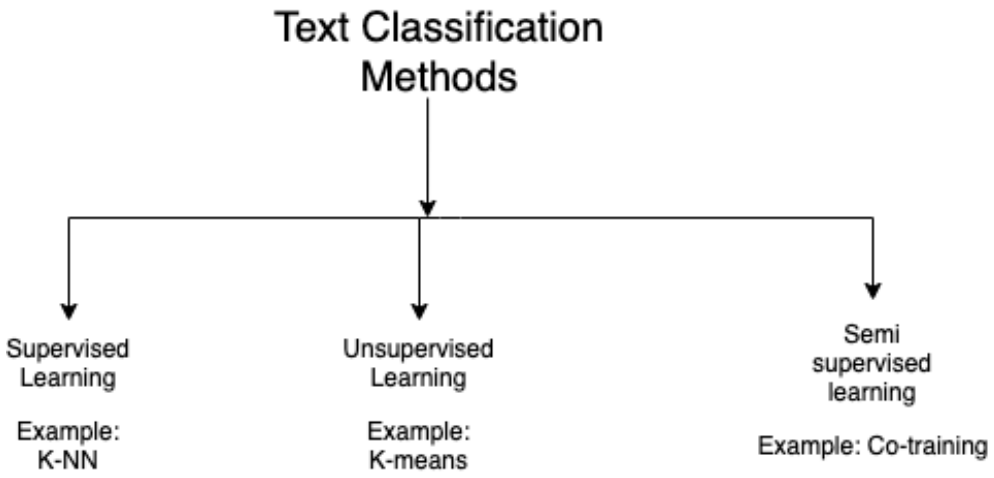

Figure 1: Text classification Methods

In this research, the texts classification follows hand-crafted classification rules, where model learn from the manually labeled training documents and categorizing the rest of the web contents. Since the text data collected from the B2B websites are massive in volume and unlabeled, classification rules were constructed and applied to manually label the training data. The process in which the classification rules were set and how the training data was verified is described in the next chapter.

In Natural Language Processing (NLP), text classification assigns class values to set of documents or Boolean values to each pair of documents and classes. For example, if

[Ci]: Predefined sets of classes, where [i = 1, 2, 3, ……., n]

[Di]: Text Documents to be classified, where [i = 1, 2, 3,……., n]

In general terms, text classification can be defined as [Di] → [Ci] where "i" can be from "1, 2, 3, ……., n". The same document can belong to one or more classes, and such a classification task is called a soft classification task. For this thesis, one document strictly belongs to one class, and this is referred to as a single label classification task.

3.1 Experiment Design

An experiment design was chosen to explain the flow of text classification task carried out. To overcome the issue of lack of labeled data to train the models, the experiment was designed to achieve the best possible accuracy in the manually annotated data through iterations.

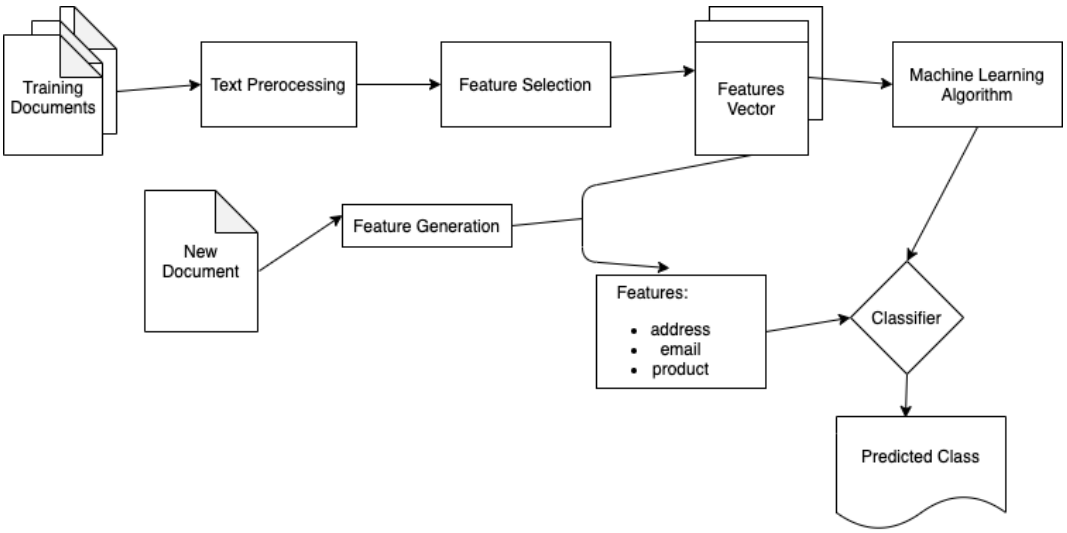

Figure 2: Framework for classification task

3.2 Training Data

Training data are essential to any automatic classification models as the machine learning algorithms learn how to make accurate predictions from the training data. Training data are labeled data used for fitting the parameters of the machine learning models. Depending on the classification tasks, training data can be labeled pictures, videos, or, like in our case, labeled text. Another concept that goes side-by-side with the training data is the testing data. Testing data is used for validating the predictive performance of the models that learned classification rule from the labeled training data. While using training data to build a machine learning model, it might be necessary to preprocess it to manage the feature space and improve the predictive performance of the model. [16]

3.3 Pre-processing

To make the text documents analyzable and classifiable, the text documents scrapped from the web has to go through preprocessing steps. Preprocessing the documents help in reducing the noise in the data and thus improve the performance of classification algorithms in most of the datasets. The noises can be the HTML tags extracted while scraping the text, stopwords, misspellings, or slang that affect the performance of the classification model. In this section, some of the standard preprocessing techniques are briefly discussed. [21] [22]

3.3.1 Punctuation

The scraped text has different types of punctuations. Punctuations are necessary for the texts' structure but do not make any contributions to the NLP model. Depending on the word embedding models we choose, we need to choose a list of punctuations that should be removed carefully. We can use python's string function to remove the following sets of punctuations:

```
import string

string.punctuation
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

In addition to these, there is also the possibility to add or remove symbols if necessary. We have to consider removing extra symbols if the performance of the model is affected.

3.3.2 Stop Words

Stopwords are prevalent and appear in large numbers in text documents. Some of the stopwords are a, an, the, is, you, etc. These words do not identify the class of the text as they frequently appear throughout the texts. NLTK provides a list of common stop words in English language and removing them helps in improving performance of text classification algorithms.

3.3.3 Tokenization

Tokenization is a technique of breaking text into small meaningful elements known as tokens. The input text is broken into tokens by reading the whitespace by default. It is also possible to customize the tokenization process

by using regular expressions. This way, developers will have more control over the output. For example,

> **Input text:** “Hello, I am a text classifier.” has the following output tokens.
>
> **Output text:** 'Hello', 'I', 'am', 'a', 'text', 'classifier'

3.3.4 Stemming

Stemming is a process of transforming words with roughly the same semantics into one standard/root form. This process helps in keeping the vocabulary small and thus, results in improved performance of many modeling tasks. Stemmers are language-specific and require less knowledge, meaning they operate on a word-by-word basis without understanding the context in which the word is used. We can change the past tense to present, for example, ate to eat, and plural to present like eggs to egg. In addition, the words computation, compute, and computers are stemmed into “comput”.

3.3.5 Lemmatization

Lemmatization works in a similar manner like stemming. The main difference is that lemmatization considers the context of the word before chopping it to its root form. Lemmatization usually takes more time than stemming as the root form generated from lemmatization are based on dictionary forms. For

example, stemmer changes the world believe in believing while lemmatization will result in the root word belief.

3.4 Feature Selection

Features define the individual measurable property of the class membership process. While there are tens of thousands of features to describe the particular process, not all are significant. Thus, selecting the features that best fits the process has been an enormous field of research in machine learning field. Feature selection or variable elimination helps in better understanding data, reducing the use of the resource, and accelerating the model execution time. The main focus of feature selection is finding the correct subset of features that best encapsulates the relationship of the individual items present in the documents. Feature selection methods are broadly classified into filter and wrapper methods. [23]

3.4.1 Filter methods

Filter methods use ranking as the primary criteria to select or discard the features. To rank the features, a ranking criterion is used, and the features that rank below the set threshold are discarded. For the features to rank higher, it has to have unique and useful information describing the item. One of the major drawbacks of this method is the issue of relevancy. The definition of unique might always not be clear, thus causing a loss of useful features. [24]

3.4.2 Wrapper Methods

In wrapper methods, the performance of the predictor is used as the basis of selecting features. Using predictors as a black box, several combinations of features are tested to get higher accuracy. The use of algorithms can combine the features, and by the use of tree structure, the subset with the best accuracy is selected. The major drawback of this method is the number of possible combinations of features. As the size of a document grows, the number of resources necessary to find the best possible combination of features can take resources and time. [24]

3.5 Text Feature representation

In this chapter, some of the popular methods used in representing the text data while modeling the text for the machine learning algorithm are discussed. There are several ways in which the features of text can be represented.

3.5.1 Bag of words

Bag of words (BOW) is one of the most commonly used and simplest text feature representation methods [46]. The core idea behind BOW is that each sentence is represented as a collection of individual word in a bag. BOW can be very useful in constructing bigrams instead of considering a single word. One of the major drawbacks of BOW is that the model does not retain any context or the order of words in the sentence.

3.5.2 Term-Frequency Inverse Document Frequency (TF-IDF)

TF-IDF is a numerical representation of how much a word is important to the document relative to the text corpus. The first part, term frequency, measures how frequently a word appears in the document and the second part, inverse document frequency, calculates the importance of the term in the document relative to the corpus. TF-IDF usually gives better performance than BOW in the machine learning models [46].

3.5.3 Word2Vec

Mikolov et al. introduced the word2vec model as an improved word embedding method, which uses a neural network with two hidden layers in order to create a higher dimension for each word vectors [49]. Many pretrained word2Vec models such as Google's pretrained model that has been trained using neural networks on news datasets are available to use on machine learning tasks. One of the major advantages of this model is that, it retains the semantic meaning of words in the documents.

3.6 Text classifiers

This chapter will present some of the most commonly used text classification algorithms for multiclass classification tasks. These algorithms are believed to be the most accurate in predicting the classes for this domain [25].

3.6.1 Support Vector Machine

Support Vector Machine or (SVM) is a supervised classification algorithm that classifies data into two available classes in its simplest form, but works in multiclass classification tasks by dividing the whole task into many binary classification tasks [26]. SVM works by finding the maximal margin separating the hyperplane between the classes in the data by constructing a hyperplane in higher-dimensional space. The output of SVM is a map of the sorted data with the distance as far as possible. Once SVM is trained with series of data, its task is to classify new data into one of the two groups. [25]

SVM simply are coordinates of individual observation and SVM classifier, which will separate the two classes by a hyperplane as shown in the picture.

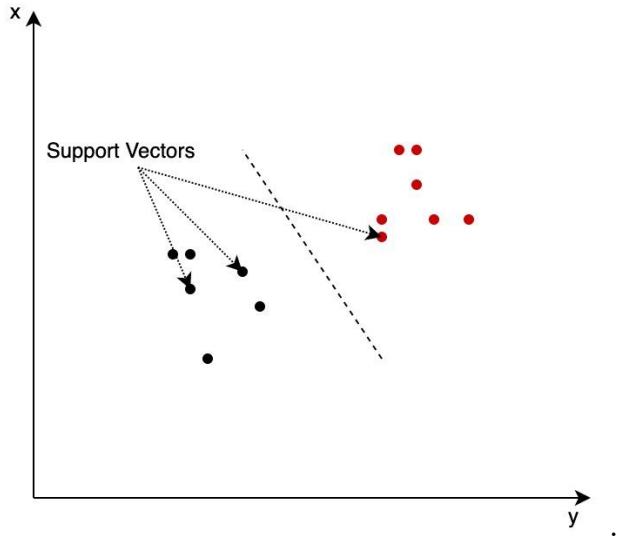

Figure 3: Decision plane in SVM

Given that already classified data is plotted in the plane, the input pair (x,y) is classified by SVM as either black or red in our example hyperplane above. The hyperplane is the decision boundary, and it is the one that maximizes the margin of both classes. Similarly, we can choose a nonlinear hyperplane for nonlinear data by adding a 3rd dimension and creating the decision boundary in a 3-dimensional hyperplane.

In order to classify text data using SVM, we need to first transfer the text into a vector of numbers. The next step is to select the feature to use, and most commonly, we use word frequencies. In other word, we treat the text data as a bag of words and for each word appearing in a specific bag we get a feature. The value of the feature will depend on the frequency of the word in the specific bag. This way, we represent our text data with thousands or tens of thousands of vector space. The next step is then to select the optimal kernel function. Every text classification problem is different, meaning we must carefully choose the kernel function to get the best possible result. [27]

Once the optimal hyperplane is selected, we change our labeled data to feature vector, train the model, and apply it to unlabeled text data to complete the classification task.

Although SVM was originally designed to be a binary classification algorithm, it dominates the multiclass classification researches now. It is one of the commonly used text classification techniques and has reliable performance. [48]

3.6.2 K-Nearest Neighbours (KNN)

K-Nearest Neighbours is a simple, easy-to-use, and low-resource-consuming algorithm that is widely used in classification and predictive tasks [58]. It is based on the idea that neighbors are of similar behavior thus, can be treated as they belong to the same group. In order to place the objects into the same group, KNN calculates the proximity of the data points. The way to calculate the proximity between the data points varies according to the task, but the most common way is to calculate the Euclidean distance or the straight-line distance. The text data should be represented in numerical form to feed it to the algorithm and thus make the classification. In order to represent the text in numerical form, we can use word embedding techniques such as Term frequency – Inverse document frequency (tf-idf), which has been discussed in detail in the above chapter, 3.5.2. [58]

The KNN classification assumes that a data instance is most similar to another data instance that is nearest or closest to it. It makes predictions based on the k similar training pattern for a given sets of new data instances, by assuming the patterns that are similar are likely of similar type. It finds a group of K objects from the training data which are closest to the test document and assigns the class. The main approach is to calculate the distance between unlabeled document to other labeled documents and find the nearest neighbors. One of the major drawbacks of this algorithm is that, it does not generate a model instead relies on the training data during the classification task. [58]

3.6.3 Random Forest

Random forest classifier is another most used classifier as it produces excellent results most of the time, even without hyper-parameter tuning [25]. Random forest classifiers are popular because they are simple to build and can be used in classification and regression tasks. Random forest is a supervised learning algorithm that works by combining learning models known as the bagging method. In other words, random forest works by building multiple decision trees and combining them to get the predictions. Random forest classifier randomly selects features to build a decision tree and combines them to get the result [29]. For example, let us consider a training set of four documents [A1, A2, A3, A4] with corresponding classes [C1, C2, C3, C4];

In such cases, the random forest may create three different decision trees making subsets of:

[A1, A2, A3],

[A2, A3, A4],

[A1, A3, A4]

Finally, the classifier will take the majority of votes from each of the decision trees made. This helps in reducing the noise that the single decision tree is prone to, thus improving the accuracy of the classification task. One of the significant advantages of using random forest classifiers is that, it is easy to build and tends to give more accurate result. This is based on the idea that when several weak estimators are combined, they seem to produce a robust estimation. Several decision trees built, few might contain some noise, but it

will not affect the overall result that much. In general, to find a better performing model than the random forest classifiers, one should spend more computational power and time. Random forest classifiers also help in overcoming the overfitting problem, given that there are enough branches present. [25]

While there are several advantages of this model, it has limitations too. Sometimes, the presence of a large number of decision trees can slow down the execution of the classifier. The training of random forest classifiers is generally faster than the actual prediction task itself. So, in the cases where the speed matters, other classifiers should be considered. Similarly, if the purpose of the classification task is to find the description of the relationship in the data, the random forest model is not suitable. [25]

3.6.4 Multinomial Logistic Regression

To provide native support for multiclass classification in Logistic Regression, multinomial logistic regression was introduced. Like the Support vector machine, Logistic Regression by default is a binary classification algorithm but can be used in a multiclass problem by dividing the whole process into multiple binary classification tasks. Multinomial logistic regression, like logistic regression, uses maximum likelihood estimation to evaluate the probability of a document belonging to a particular class. [30]

In order to get from a binary to multiclass classification with “n” numbers of classes, the multinomial regression model assumes an “n-1” numbers of logit equations. The model compares all the combinations of “n” groups. Logit

equations are logarithmic functions to restrict the values between 0 and 1. At the center of this multinomial logistic regression classification task there lies the task of estimating the log odds of each class. The logit probability can be explained by the following equation:

$$Logit(X(p)=\ln((X(p=1|s))/(1-X(s=1|p)))$$

So, given some feature p, the model tries to calculate the probability of event "s". So, "s" can either be 0 or 1. For example, "y" can be whether a particular document belongs to a class or not. If it belongs to the class, value "1" can be assigned, and if it does not, then value "0" is assigned.

3.6.5 Hybrid classifiers

Hybrid classifiers are classification techniques that uses multiple classifiers that act in compliment of each other. Each method has a different classification task, and when all the tasks are completed, the decision is made by one method. One data instance that might be misclassified by a classifier may be correctly predicted by another, thus reducing the number of misclassified data instances.

3.6.6 Neural Network Classifiers

A neural network classifier consists of neurons (units) arranged in a layer to convert the input vector into the output. In a multiclass classification task, the

number of output options is the same as the available classes. Each output node belongs to a class, and the output is the score relating to that class. A hidden layer calculates the score between the inputs and the outputs. One of the significant advantages of neural network classifiers is that it allows building a very simple to complex architecture as it is made up of simple blocks. [31]

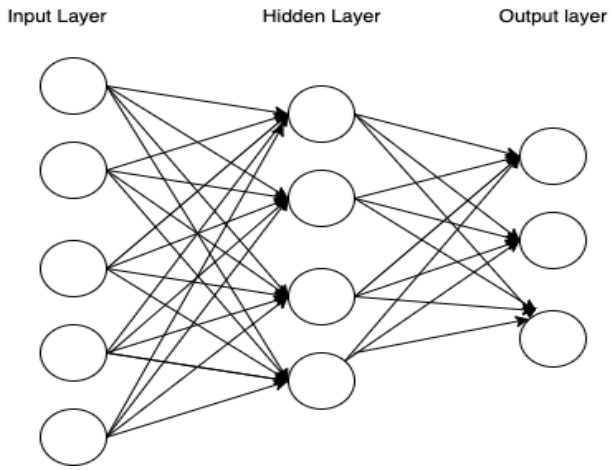

Figure 4. Simple architecture of a neural network classifier

3.6.7 Gradient Boosting

Boosting method works by combing a set of weak classifiers to deliver efficient and improved classification results. For any data instance "i", the weight of the model instance is assigned based on the previous instance "i-1", and the classified outcome that is correct is given lower weight while the wrongly classified instance is given more weight. This way the model focuses on misclassified data instances and predicts them correctly. Several iterations of the weighing are performed on different models and in the end delivering

a more consolidated result. The attributes tested for gradient boosting are learning rate, maximum depth, max features, and minimum samples split. [57]

For a training dataset, $A = [I_i, N_i]$ the aim of the gradient boosting model is to map the data instances "I" to the outputs "N" by minimizing the expected value of the loss function and increasing the accuracy.

3.7 Performance Evaluation

In order to find out that the classification algorithms learned something, it is essential to calculate the performance. The following classification metrics help in evaluating the performance of the classifiers.

a. Precision [tp/ (tp+fp)]: Precision measures the ability of a classification algorithm to identify only the instances belonging to a particular class successfully.

b. Recall [tp/ (tp+fn)]: Recall measures the ability of a classification algorithm to identify all the instances belonging to a particular class successfully.

c. F1 score: F1 score is the weighted mean of precision and recalls normalized between 0 and 1. The score 1 for F indicates the perfect balance between recall and precision as they are inversely proportionate.

d. Support: Support is the actual occurrence of the class in the test dataset. If the support is imbalanced, it might signal the need to update the training dataset to achieve a higher balance.

3.7.1 Confusion Matrix

Confusion matrix plot the combination of actual and predicted classes. Each row in the matrix denotes the data instances in the predicted class, while each column represents the data instances in the actual class. The confusion matrix shows weather models can consider the overlap on the properties and which classes cause the most confusion, thus driving the overall accuracy of the prediction.

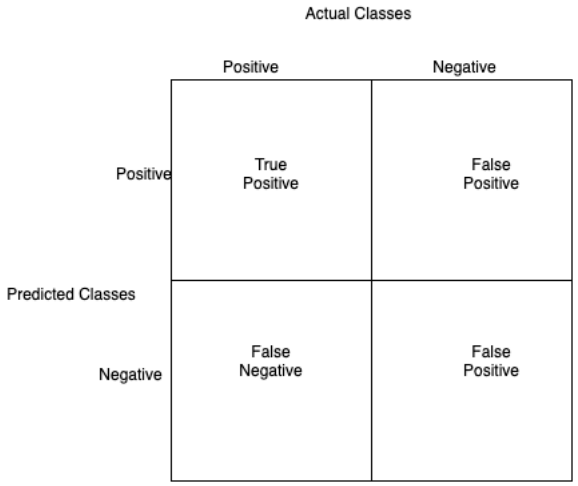

Figure 5: Confusion Matrix

If the classification algorithm correctly predicts the instance to a class, it is referred to as true positive, but if a negative instance is predicted, it is referred to as true negative. However, if the algorithm wrongly predicts the data instance to be positive when in actuality, it is negative, it is referred to as a

false positive, and vice versa is a false negative. The classification results are commonly visualized using a confusion matrix, as shown in the picture above.

4 Training Data Creation

4.1 Dataset

The data used in this study was obtained from a case company that used a third party, a leading account-based marketing platform in Europe, to collect the data. As a part of data collection, the case company had an agreement with B2B companies to collect cookies-based data from the registered companies. The data used for this thesis is a part of that collected data from 74 different B2B companies from different service sectors like automotive, IT services, and health care. The dataset was massive and contained browsing sessions from about 180 million browsing history. For this project, the text contents that the browsers were viewing during the sessions are used. Among the pages, about 60% were from the seller's channels, and the rest of the 40% were promotional channels that contained content from the sellers. The size of the data made available for this research was about 2.3GB in size.

A	B	C	D	E	F	G
	Unnamed: 0	url_id	content	lang	domain	site_info_url_id_x
0	0	4394953	\tOWNERS of new-build flats in Bournemouth say..	en	NaN	4394953.0
1	1	21005737	\tBrooks Agnew has faithfully served NPIEE tea...	en	www.abovetopsecret.com	21005737.0
2	2	21005737	\tBrooks Agnew has faithfully served NPIEE tea...	en	www.abovetopsecret.com	21005737.0
3	3	4162146	\tWASHINGTON (Reuters) - A grand jury has issu...	en	NaN	4162146.0
4	4	10711309	\tGlenn McCoy Belleville News-Democrat\n\tLee ...	en	NaN	10711309.0

Figure 6: sample of data

In this research, only the contents in English were used.

	URLID	CONTENT	CLASS
0	20486341	\t1 October 2017 Excluding LNG carriers and inland waterway vessels Source: Courtesy of DNV GL LNGi, Excluding LNG carriers and inland waterway vessels © Wärtsilä 4 DNV GL LNGI - AREA OF OPERATION, OF LNG FUELLED VESSELS 4 Updated 1 October 2017 Excluding LNG carriers and inland waterway vessels Area, waterway vessels are also shown, but these projects are not counted here. 1 2 3 1 1 Within end, LNGi portal (dnvgl.com/lngi) Updated 1 October 2017 *In the LNGi map bunkering barges for inland <some link> Interim report Q4 and full year 2012	Product/Service
1	4377464	285 \treplacement and speed / load controller upgrade Wärtsilä Ecometer - automation upgrade solution, - Automatic Voltage Regulators retrofit solutions Field services (E&A) Governor <some link> Governor replacement	Product/Service
2	15223535	\tseals, in-water serviceable seal, Sandguard , We offer both face and lip seals to deal with a <some link>	Product/Service
3	5242983	\tGate Valves - Wafer - Wärtsilä Valves, valve - frontview Gate Valves - Wafer Gate Valves - Wafer, Gate valves , wärtsilä gate valve , gate valve wafer, wafer valve, valve wafer, Wärtsilä Gate Valves - wafer, available in Carbon Steel, Stainless Steel & Ferralium®. , Gate Valves - Wafer Size Range 6" - 72, Gate - Wafer - C G9 gate , Duplex & Super Duplex Valves Download datasheet C-G9 Wafer Gate <some link> Pumps & Valves	Product/Service
4	4960051	\t, W16V34DF , W46DF (L - version) TCH380 3800 W8V31, W10V31 Double Input Gears 109 Propulsors & Gears 2 <some link>	Product/Service

Figure 7: sample of English data only

4.2 Pre-processing

This pre-processing of data was conducted using python 3.7, a popular programming language in the field of NLP. The collected text documents were of different languages so the first step was to filter the contents to extract documents of English language only, in which the experiment was conducted. Out of the total dataset, 39% were in English. Python's langdetect package was used to filter out the non-English contents. The extracted English content data was then saved as a different subset.

4.3 Training Data

The training data was labeled manually from the available dataset for this experiment. The first step in creating the training data was defining the classes. The class definition was done in several iterations. First, using the K-means clustering and elbow method, the entire content of the data was plotted to see the optimal number of clusters [28]. This gave a sense of how many classes we should be looking at. In order to better understand how the classes should be generated and how the actual content is developed for the B2B websites, Russel Mattinson from Stora Enso was interviewed, who works as Head of Digital Engagement at Stora Enso. During the one and half-hour-long interview, Mr. Mattinson gave an in-depth process of how the contents are developed for which part of the digital platform. One of the main focuses was to make the customer come back to get more details once they start looking for the product, which means giving the right amount of information at the correct time. His input was constructive while defining the classes.

From the English subset of the data, the training data was sampled. Once the idea behind creating the digital content to target a specific audience group were gathered, the next step was to use that knowledge to classify available documents into those respective classes. The content from Stora Enso was filtered first in the available dataset by matching the domain name using regular expression in python. Unfortunately, the amount of data available from the Stora Enso domain was not enough to create the training data; hence the next most extensive domain set, Wartsila, was also selected. Selecting Wartsila and Stora Enso's available data gave 15120 documents to create training data. The next step was going through each row and manually labeling the data based on the classification rules created.

During the manual class coding process, it was soon clear that it would be an imbalanced training dataset. There were incomplete and unusable data and also not enough data belonging to the class contacts and careers; hence, only Maximum of 200 documents belonging to one class were manually labeled. To make the dataset more or less balanced, some data belonging to the contact class was manually added.

Once the desired subset was created, the next step was to apply the cleaning and preprocessing steps. The first step in the data cleaning involved removing the hyperlinks, HTML tags, and pointers to the images. After the first iteration of the training data, it was realized that on a significant portion of the collected document, the footer explaining the stock listing and

operation of the company was also present. As this is present in all different kinds of pages, those texts would affect the classification task; hence such a portion was selected and removed using regular expression in python. The next step was to remove the stop words, convert all the texts to lower case, and tokenize the words using whitespace.

By interpreting the result and discussing the need to use this research for further studies and other ongoing parallel projects, it was decided that the text documents will be divided into seven classes. Below is the definition of the class, what it is for this particular dataset, what it is not, and a part of text extract as an example. This exact definition was used to verify the class definition during the cross-developer verification, which is discussed in detail in the training data verification part.

The following table summarizes the classes, their definitions and contains an example, from the dataset used in the research.

S.N.	Class	What it is	What it is not	Example
1.	Product/ Service:	The documents belonging to this class present the item or items that are manufactured for sale and have the lists of actions performed according to customers' demands and where no physical goods are transferred. This class contains a general idea of the items or services provided by the company for the customer to get a general idea. The idea behind the text contents in this class is to make sure that the customers get a general overview of the products and overwhelm them with details. It is expected that customers will scan through the information and decide if they need to look for more details or keep looking. Generally speaking, the text of the documents in this class is kept as little as possible.	This class does not list the detailed physical or technical specifications of the product and has no details regarding the tasks performed. The documents in this class are not meant to provide any in-depth information on the product. The information in the document belonging to this class is not meant to make the customer decide on a purchase but to push them to get more details.	*: \treplacement and speed / load controller upgrade Wärtsilä Ecometer - auto mation upgr. ade solution, - Automatic Voltage Regul ators retrofit solutions Fie ld services (E&A) Govern or <some link> Governor replacement*
2.	Detailed Product/ Service	The document belonging to this class presents a detailed description of any product or service provided by the company. Details Include product specifications, technical	It is not the detailed description of the product or service described during any announcement made by the company while publishing news regarding won bids or	*Example: \tSpeed 14.0 kno ts LNG cargo capacity 10, 000 m3 MDO/ MGO stora ge capacity 2,400 m3 Gas,) and MDO/ MGO outside at competitive fuel costs R*

		specifications, and procedures on how the tasks will be completed and how the product operates. The document in this class should have enough information for the customer to decide whether it is the solution they are looking for. The main idea for the content on this page is to help the customer make the purchase decision and make initial contact with the agent.	service agreement signed. While putting out announcements regarding the contracts won or projects completed, companies tend to provide some information regarding what will be built. That information is targeted to a different group of visitors than the potential customers.	*educed emissions in LNG operation: SOx (100, carg o tanks Abt. 2,400 m3 of c argo capacity for marine d iesel oil (MDO/ MGO) C ombined manifold for, LN G or MDO/ MGO cargo/f uel operations <some link > MV Theben*
3.	General Announ cement	The documents in this particular class present publications by the company for anyone interested in the company. It includes what is happening in the company; bids won, annual reports published, and upcoming events. The announcement regarding bids won may have some details about the service being provided and products being used.	The documents in this class do not provide the financial details of the deals in details, and do not have the details of the annual financials of the company.	*\tWärtsilä Corporation's Annual Report 2015 publi shed, Wärtsilä Corporatio n has today, 9 February 2 016, published its annual r eport for the year 2015 on , The electronic annual re port contains the Business review, the complete, Fina ncial Statements 2015. Th e annual report also inclu des a Wärtsilä Stories sect ion with further, over 200 locations in more than 70 countries around the worl d. Wärtsilä is listed on Na sdaq Helsinki Wärtsilä An nual Report 2015 (pdf), in formation on Wärtsilä's b usiness environment and o n sustainability. A PDF-fil e of the annual*
4.	Detailed Announ cement	The documents in this class provid e detailed financial information fo r stakeholders in the company who have more investments in the com pany. These documents are also in tended for potential customers wh o want to know their financial hea lth before making any agreements. All the financial and structural inf ormation that could be made publi c are generally listed in documents in this category.	It is not an announcement that has details about the new operation being started with the details about the capacity of the new operations.	*: \tWärtsilä's acquisition i n Singapore has been app roved by the shareholders of Total Automation Ltd in an Extraordinary General Meeting today. In Februar y Wärtsilä signed an agre ement to acquire the entir e business of Total Autom ation Ltd and all the share s in its subsidiaries. The tr ansaction is expected to cl ose within the second quar ter of 2006.\n\tThe transac tion price is EUR 61.6 mil lion. Total Automation's n et sales in 2005 totalled E UR 42.8 million and it has consistently generated EB IT margins in excess of 15 %. The company employs over 400 people and has f*

				acilities in Singapore, Dub ai, France, the UK and Ch ina.\n\tFurther informatio n: Tage Blomberg, Group Vice President, Wärtsilä S ervices, phone +358 10 70 9 2425
5.	Careers	The traffic in the B2B websites might not only be by the potential customers but also by potential employees. The contents of the documents in this class present stories about working/have worked in the company to promote work-life at the company, messages about the trainees, and vacancy announcements. The motivational stories to attract more applications and the best candidates for any positions are presented in the documents of this class.	It is not the announcement of a change in leadership roles in a company. Such announcements are primarily designed for different audiences. For example, the following content by Wartsila is primarily for anyone who wants to know about the company.	*\tgot to work as a trainee i n Wärtsilä's Nuclear EDG Projects. My job was to as sist the project teams, easy decision to apply again fo r a summer internship. La st summer I worked in Nuc lear EDG projects as*
6.	Contact	The contents of documents in this class list the contact details of the office branches, contact details of company representatives. Once the customer decides to approach, they should find the information regarding who and where they should approach.	It is not a contact detail listed in the texts published as news or announcements by the company.	*tDr. Joe Thomas Director, Ballast Water Managemen t Systems, Environmental Solutions Wärtsilä Marine Solutions Tel: +44 1202 6 62600 joe.thomas@wartsi la.com\n\tHanna Viita Dir ector, Marketing Wärtsilä Services hanna.viita@war tsila.com Tel: +358 40 16 7 1755*

Table 1: Classes, their definition, and examples

Based on these class definitions a set of training data was created with the following distribution.

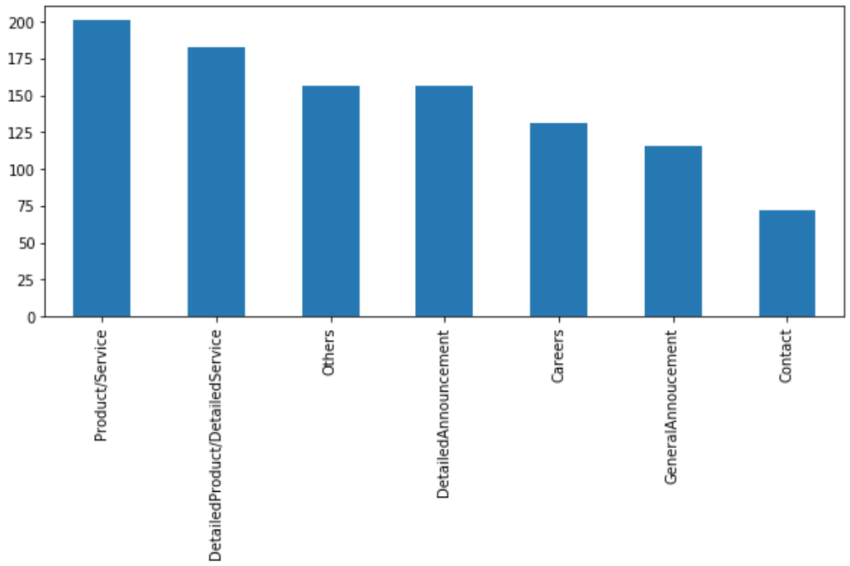

Figure 8: Distribution per class of the manually labelled data

To get focus more on the research objective and get more balanced training data distribution, careers and contact classes were also merged to the others making it five-class classification tasks. The final shape of the training data is presented in the picture below.

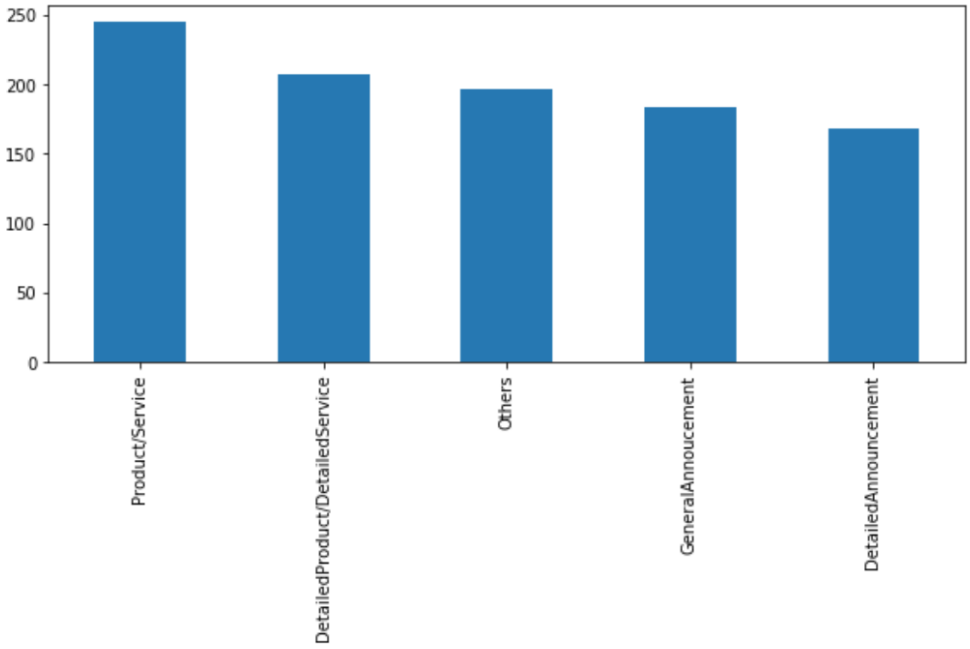

Figure 9: Distribution per class of manually labeled data used for the training.

4.4 Training data Verification

The initial plan to construct the class definition was by interviewing a few digital content creators. Emails were sent out trying to set up the interviews but not enough interviews were set-up. The one interview that was conducted helped the research to define the classes but was not enough to verify the definitions. Hence to overcome that problem, a cross-developer verification was carried out. For this purpose, five members of this project were provided with the class definitions and documents they needed to classify manually based on those rules. The idea was to have at least 80% agreement between classes manually labeled by all participants. During the first iteration, there was less than 80% agreement. Hence the definition was improved to contain what is not part and the examples. This helped in improving the understanding between the members, and the required level of agreement was achieved.

Once the class definitions were agreed upon, the next step was to classify the text documents randomly selected from the dataset manually.

5 Experiment and Results

As described in the chapters before, automatic text classification is widely adopted across many domains, and it helps organize the documents and gain valuable insights. As a part of text classification, web analytics is gaining popularity as many documents are produced and distributed digitally at present. The number of documents available digitally keeps growing without an accurate idea of what can be achieved by analyzing those. Explicitly speaking, not much research was found in the field of B2B text document classification. This opens up a completely new field of research; hence, one of the main challenges was to define the scope of this Master's thesis. Due to the lack of training data available, much time was spent in creating the training data manually. It is no secret that the analyzing the available text data and combining the research with the time spent and times visited even more valuable insights could be obtained.

After using much of the time to create the training data through different iterations, time was of constraint; the scope of the research had to be limited to training few popular text classification models using the training data created and interpreting those results.

5.1 Resources

5.1.1 Business Finland

Business Finland is a government organization in Finland promoting innovation, funding and trade, and travel. Business Finland is a part of the Team Finland network. As a project of implementing artificial intelligence in promoting the B2B sales, Business Finland has used N. Rich to collect various datasets from companies inside Finland and made them available for the project. This kind of project aims to provide an opportunity to use insights and promote the companies going global. For this project, Business Finland provided data continuously and helped in defining the problems.

5.1.2 Aalto University, School of Business

Aalto University, as a general, is one of the tops, if not the top, universities in Finland. School of Business, in particular, is focused on working on better business and a better society. Through collaborations with other branches and programs like Information Systems management, the business has managed projects concerning data analytics and artificial intelligence. Being the project owner, Aalto University provided access to the data, necessary tools like computers and workspace, and a supervisor who constantly helped throughout the research.

5.1.3 Compute Resources and Tools

Compute resource for conducting this project was provided by Aalto University. Access to Aalto's Trition cluster was provided. Trition is Aalto's high-performing computing cluster.

The experiment was conducted using python 3.7, a popular programming language in natural language processing. The notebooks were executed on the triton cluster.

5.2 Results

The preprocessed, cleaned, and manually labeled training data was used to extract a matrix of TF-IDF features. The training data was then split into sets of the train, and test data with the ratio of 4:1, meaning the shape of the train set was 849, and the test data shape was 150 with 4804 TF-IDF features computed in total. The next important step was to find the algorithm that performs best in this particular domain. Selected algorithms described in the theoretical section were used to train the models. The selection was based on researches presented in Chapter 2; Related works. The trained models were saved and later used to classify the 372550 unlabeled text documents.

5.2.1 Multinomial Logistic Regression

Logistic Regression was one of the algorithms used for this classification task. First, a model was built with random parameters to find the best hyperparameters. On the random search, Logistic Regression was able to predict the classes with an accuracy of 0.75. Logistic regression performed better when fine-tuned after searching for the best parameters. On grid search the accuracy of the model increased to 0.81.

On grid search the training accuracy of logistic regression was more than 0.94 and the test accuracy was 0.81. As we can see from the classification report below, the class Detailed Product had the lowest accuracy causing the decrease in accuracy of the model.

CLASSIFICATION REPORT

	PRECISION	RECALL	F1-SCORE	SUPPORT
PRODUCT/SERVICE	**0.80**	**0.76**	**0.78**	**37**
DETAILEDPRODUCT/ DETAILEDSERVICE	**0.65**	**0.71**	**0.68**	**31**
GENERALANNOUCEMENT.	**0.83**	**0.86**	**0.84**	**28**
DETAILEDANNOUCEMENT	**0.95**	**0.76**	**0.84**	**25**
OTHERS	**0.91**	**1.00**	**0.95**	**29**
ACCURACY			**0.81**	**150**
MACRO AVG	**0.83**	**0.82**	**0.82**	**150**
WEIGHTED AVG	**0.82**	**0.81**	**0.81**	**150**

Table 2: Classification report of logistic regression on training data

5.2.2 Linear Support Vector Machine

Linear Support Vector Machine (LSVM) was the following algorithm to be tested in this dataset. First a random search was performed to find the best hyperparameters for fine tuning the model. On random search LSVM was able to predict the classes with the mean accuracy of 0.75.

On finetuning, the accuracy of LSVM was up to 0.79. The training accuracy of LSVM was 0.97 while the test accuracy was 0.79. On the classification report below, we can see that the mean accuracy was affected again by class “Detailed Product” as it had the lowest accuracy rate amongst the classes.

CLASSIFICATION REPORT

	PRECISION	RECALL	F1-SCORE	SUPPORT
PRODUCT/SERVICE	0.76	0.78	0.77	37
DETAILEDPRODUCT/ DETAILEDSERVICE	0.66	0.68	0.67	31
GENERALANNOUCEMENT.	0.79	0.79	0.79	28
DETAILEDANNOUCEMENT	0.95	0.72	0.82	25
OTHERS	0.88	1.00	0.94	29
ACCURACY			0.79	150
MACRO AVG	0.81	0.79	0.80	150
WEIGHTED AVG	0.80	0.79	0.79	150

Table 3: Classification report of LSVM on the training dataset

5.2.3 Random forests

A similar approach to the above two algorithms was taken for this classifier as well. First, the default hyperparameters of the model were searched, and the grid was defined, after which a random search was performed. On random search, Random Forest was able to predict the classes with a mean accuracy of 0.76. From the random search, we learned the best hyperparameters, and more in-depth fine-tuning was performed around those hyperparameters. On finetuning the mean accuracy of the Random Forest classifier was up to 0.79.

The picture below shows that the random forest model struggled in the class products/service and detailed products/services and did well in the rest of the classes

CLASSIFICATION REPORT

	PRECISION	RECALL	F1-SCORE	SUPPORT
PRODUCT/SERVICE	0.68	0.86	0.76	37
DETAILEDPRODUCT/ DETAILEDSERVICE	0.66	0.61	0.63	31
GENERALANNOUCEMENT.	0.91	0.75	0.82	28
DETAILEDANNOUCEMENT	0.95	0.80	0.87	25
OTHERS	0.90	0.93	0.92	29
ACCURACY			0.79	150
MACRO AVG	0.82	0.79	0.80	150
WEIGHTED AVG	0.81	0.79	0.79	150

Table 4: Classification report of Random Forest on the training data

5.2.4 K-Nearest Neighbor (KNN)

KNN performed poorly in this particular dataset. On random search, the mean accuracy of KNN was 0.70. On fine-tuning, the model during the grid search the mean accuracy could be increased to 0.74. Even on the training set, the accuracy of KNN was 0.73. As seen from the picture below KNN struggled to predict classes Product/Service, Detailed Product/Detailed Service, and General Announcements.

CLASSIFICATION REPORT

	PRECISION	RECALL	F1-SCORE	SUPPORT
PRODUCT/SERVICE	**0.66**	**0.57**	**0.61**	**37**
DETAILEDPRODUCT/ DETAILEDSERVICE	**0.66**	**0.61**	**0.63**	**31**
GENERALANNOUCEMENT.	**0.66**	**0.82**	**0.73**	**28**
DETAILEDANNOUCEMENT	**0.90**	**0.76**	**0.83**	**25**
OTHERS	**0.88**	**1.00**	**0.94**	**29**
ACCURACY			**0.74**	**150**
MACRO AVG	**0.75**	**0.75**	**0.75**	**150**
WEIGHTED AVG	**0.74**	**0.74**	**0.74**	**150**

Table 5: Classification report of KNN on training data

5.2.5 Gradient Boosting

Parameter tuning with gradient boosting was used as the last method for classification. The performance of gradient boosting was underwhelming than expected. The boosting method was able to predict the classes with an accuracy of 0.76. The complete classification report can be seen in the table below.

CLASSIFICATION REPORT

	PRECISION	RECALL	F1-SCORE	SUPPORT
PRODUCT/SERVICE	**0.66**	**0.78**	**0.72**	**37**
DETAILEDPRODUCT/ DETAILEDSERVICE	**0.79**	**0.63**	**0.70**	**30**
GENERALANNOUCEMENT.	**0.70**	**0.74**	**0.72**	**19**
DETAILEDANNOUCEMENT	**0.95**	**0.75**	**0.84**	**24**
OTHERS	**0.80**	**0.86**	**0.83**	**37**
ACCURACY			**0.76**	**147**
MACRO AVG	**0.78**	**0.75**	**0.76**	**147**
WEIGHTED AVG	**0.77**	**0.76**	**0.76**	**147**

Table 6: Classification report of Gradient boosting on training data

5.3 Predicting Unlabeled Texts

Once the models were trained and saved then they were used to classify the remaining unlabeled texts. The unlabeled text documents went through the same exclusion criteria and text cleaning process described in the training data creation techniques. Since none of the models performed absolutely in random search or grid search, a voting method was introduced to get the final class of each document. The 373550 documents were classified using each model and the class that had the highest vote from each model was then used as the assigned class to the document. KNN model had performed poorly on the training data set and in the final vote caused roughly 19 000 documents to have equal votes. Hence, the model KNN random search was excluded from the final voting model. After that, the number of documents having equal votes for two classes was down to 190. The final class distribution after voting is presented in the picture below.

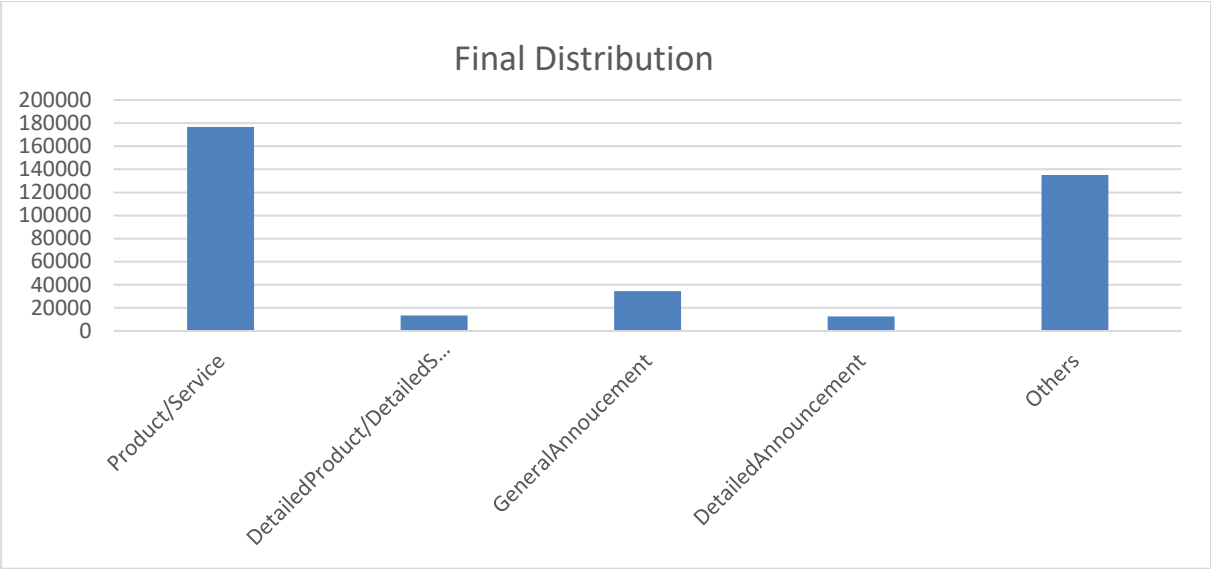

Figure 10: The class distribution of the unlabeled documents

As seen in the figure above, the majority of the digital B2B web content belongs to the class Product/Service, others being the second biggest category. The others class contained documents related to careers, support, and contacts as well, so the result is not a surprising one. The number of documents from class General Announcements and Detailed Announcements was the lowest while detailed products had just above 25 000 documents. Since the documents were mostly collected from customer sessions the result is within the expected scope. The pages showing the general products are visited most and the customer journey continues to the detailed product and contact section.

6 Discussion

This thesis presents an overview of text classification, its processes, and popular text classification algorithms. In addition to that, a research framework to classify digital contents from B2B company websites is presented. It is clear that in the field of B2B marketing, there is a severe lack of research and tools on how the large amount of freely available digital data could be utilized, and this thesis tries to bridge that gap by creating manually labeled datasets as training data. The manually labeled dataset was then used to train the text classification algorithms. The text data was first preprocessed, and then the features were extracted using the TFIDF vectorizer. Multinomial Logistic Regression was the best performing model amongst the tested models.

The results were underwhelming, but the results can provide a pathway for future results. In the B2B marketing context, classes like careers are unnecessary, although it is imperative from the organizational point of view. For the marketing research, it might be better if this class is also classified as others. In the dataset the experiment was conducted in, it is difficult to find enough training data for all the classes to make a large and unbiased training set. Also, the classes like general announcements, detailed announcements, and job announcements will have some form of contact information in them. So, for training data of this size, it was difficult for these models to predict contact class with higher accuracy. This, in turn, drove the average performance of the model largely. The prediction accuracy of the rest of the classes remains encouraging.

As a result of this thesis, the trained models were saved and the collected text classes were classified into five classes. The preprocessed document is saved hence if required can be used for classification tasks that might have more or fewer classes than this research has. The class definition of seven possible classes with examples is already created and validated although only five classes were eventually used for this thesis task. This result and models can be used as a baseline for further research, and with enough training data, the task looks promising.

For future study, data can be collected for this research purpose from which more balanced and more extensive training data could be developed. Different word embedding techniques that preserve the semantic relation of the words in the document might perform better.

References

1. Upreti, B.R., Huhtala, J.P., Tikkanen, H., Malo, P., Marvasti, N., Kaski, S., Vaniala, I. and Mattila, P., 2021. Online content match-making in B2B markets: Application of neural content modeling. Industrial Marketing Management, 93, pp.32-40.
2. Kaushik, A., 2009. Web analytics 2.0: The art of online accountability and science of customer centricity. John Wiley & Sons.
3. Rowley, J., 2008. Understanding digital content marketing. Journal of marketing management, 24(5-6), pp.517-540.
4. Sotarauta, M., Ramstedt-Sen, T., Kaisa Seppänen, S. and Kosonen, K.J., 2011. Local or digital buzz, global or national pipelines: patterns of knowledge sourcing in intelligent machinery and digital content services in Finland. European Planning Studies, 19(7), pp.1305-1330.
5. Sebastiani, F., 2005. Text categorization. In *Encyclopedia of Database Technologies and Applications* (pp. 683-687). IGI Global.
6. Bhumika, P.S.S.S. and Nayyar, P.A., 2013. A review paper on algorithms used for text classification. International Journal of Application or Innovation in Engineering & Management, 3(2), pp.90-99.
7. Attardi, G., Gullì, A. and Sebastiani, F., 1999. Automatic Web page categorization by link and context analysis. In Proceedings of THAI (Vol. 99, No. 99, pp. 105-119).
8. Jensen, R. and Shen, Q., 2007. Fuzzy-rough sets assisted attribute selection. IEEE Transactions on fuzzy systems, 15(1), pp.73-89.
9. Bodon, F., 2003, November. A fast APRIORI implementation. In FIMI (Vol. 3, p. 63).
10. Goyal, R.D., 2007, November. Knowledge based neural network for text classification. In 2007 IEEE International Conference on Granular Computing (GRC 2007) (pp. 542-542). IEEE.
11. Tasci, S. and Gungor, T., 2008, October. An evaluation of existing and new feature selection metrics in text categorization. In 2008 23rd International Symposium on Computer and Information Sciences (pp. 1-6). IEEE.
12. 12. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A. and Edwards, L., 2009. Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB, 2, pp.1-7.
13. 13. Khan, A., Baharudin, B., Lee, L.H. and Khan, K., 2010. A review of machine learning algorithms for text-documents classification. Journal of advances in information technology, 1(1), pp.4-20.
14. 14. Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
15. 15. Le, Q. and Mikolov, T., 2014, June. Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.
16. Arusada, M.D.N., Putri, N.A.S. and Alamsyah, A., 2017, May. Training data optimization strategy for multiclass text classification. In 2017 5th International Conference on Information and Communication Technology (ICoICT) (pp. 1-5). IEEE.
17. 17. Parmar, P.S., Biju, P.K., Shankar, M. and Kadiresan, N., 2018, September. Multiclass text classification and analytics for improving customer support response through different classifiers. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 538-542). IEEE.
18. 18. Rennie, J.D. and Rifkin, R., 2001. Improving multiclass text classification with the support vector machine.
19. 19. Liao, S.H., 2005. Expert system methodologies and applications—a decade review from 1995 to 2004. Expert systems with applications, 28(1), pp.93-103.
20. 20. Yu, B., 2008. An evaluation of text classification methods for literary study. Literary and Linguistic Computing, 23(3), pp.327-343.

21. 21. HaCohen-Kerner, Y., Miller, D. and Yigal, Y., 2020. The influence of preprocessing on text classification using a bag-of-words representation. PloS one, 15(5), p.e0232525.
22. 22. Vijayarani, S., Ilamathi, M.J. and Nithya, M., 2015. Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 5(1), pp.7-16.
23. 23. Kira, K. and Rendell, L.A., 1992. A practical approach to feature selection. In Machine learning proceedings 1992 (pp. 249-256). Morgan Kaufmann.
24. 24. Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering, 40(1), pp.16-28.
25. 25. Pranckevičius, T. and Marcinkevičius, V., 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing, 5(2), p.221.
26. 26. Mathur, A. and Foody, G.M., 2008. Multiclass and binary SVM classification: Implications for training and classification users. IEEE Geoscience and remote sensing letters, 5(2), pp.241-245.
27. 27. Bo, G. and Xianwu, H., 2006. SVM multi-class classification. Journal of Data Acquisition & Processing, 21(3), pp.334-339.
28. Li, Y. and Wu, H., 2012. A clustering method based on K-means algorithm. *Physics Procedia*, *25*, pp.1104-1109.
29. Prinzie, A. and Van den Poel, D., 2008. Random forests for multiclass classification: Random multinomial logit. Expert systems with Applications, 34(3), pp.1721-1732.
30. Starkweather, J. and Moske, A.K., 2011. Multinomial logistic regression.
31. Richard, M.D. and Lippmann, R.P., 1991. Neural network classifiers estimate Bayesian a posteriori probabilities. Neural computation, 3(4), pp.461-483.
32. Sokolova, M., Japkowicz, N. and Szpakowicz, S., 2006, December. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.
33. N. Rich advertisment, https://www.n.rich/
34. Business Finland, https://www.businessfinland.fi/en/for-finnish-customers/home
35. Python Packages: https://pypi.org/
36. Stora Enso: https://www.storaenso.com/en
37. Wartsila: https://www.wartsila.com/
38. Sebastiani, F., 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), pp.1-47.
39. Faraz, A., 2015. An elaboration of text categorization and automatic text classification through mathematical and graphical modelling. Computer Science & Engineering: An International Journal (CSEIJ), 5(2/3), pp.1-11.
40. Hotho, A., Nürnberger, A. and Paaß, G., 2005, May. A brief survey of text mining. In *Ldv Forum* (Vol. 20, No. 1, pp. 19-62).
41. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S. and Ré, C., 2017, November. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases* (Vol. 11, No. 3, p. 269). NIH Public Access.
42. Batista, G.E., Prati, R.C. and Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), pp.20-29.
43. Hartmann, J., Huppertz, J., Schamp, C. and Heitmann, M., 2019. Comparing automated text classification methods. International Journal of Research in Marketing, 36(1), pp.20-38.

44. Wang, L. and Zhao, X., 2012, April. Improved KNN classification algorithms research in text categorization. In 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet) (pp. 1848-1852). IEEE.
45. Grewal, D., Hulland, J., Kopalle, P.K. and Karahanna, E., 2020. The future of technology and marketing: a multidisciplinary perspective.
46. Utterback, J., 1994. Mastering the dynamics of innovation: How companies can seize opportunities in the face of technological change. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship
47. Chen, M., Weinberger, K.Q. and Sha, F., 2013. An alternative text representation to tf-idf and bag-of-words. arXiv preprint arXiv:1301.6770.
48. Weinberger, K., Dasgupta, A., Langford, J., Smola, A. and Attenberg, J., 2009, June. Feature hashing for large scale multitask learning. In Proceedings of the 26th annual international conference on machine learning (pp. 1113-1120).
49. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L. and Brown, D., 2019. Text classification algorithms: A survey. Information, 10(4), p.150.
50. Almeida, F. and Xexéo, G., 2019. Word embeddings: A survey. arXiv preprint arXiv:1901.09069.
51. Wang, W.Y.C. and Wang, Y., 2020. Analytics in the era of big data: the digital transformations and value creation in industrial marketing.
52. Lilien, G.L., 2016. The B2B knowledge gap. International Journal of Research in Marketing, 33(3), pp.543-556.
53. Wall, A. and Spinuzzi, C., 2018. The art of selling-without-selling: Understanding the genre ecologies of content marketing. Technical Communication Quarterly, 27(2), pp.137-160.
54. Dzyabura, D. and Yoganarasimhan, H., 2018. Machine learning and marketing. In Handbook of Marketing Analytics. Edward Elgar Publishing.
55. Cui, G., Wong, M.L. and Lui, H.K., 2006. Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. Management Science, 52(4), pp.597-612.
56. Ma, L. and Sun, B., 2020. Machine learning and AI in marketing–Connecting computing power to human insights. International Journal of Research in Marketing, 37(3), pp.481-504.
57. Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), pp.1937-1967.
58. Joseph, F. and Ramakrishnan, N., 2015. Text categorization using improved K nearest neighbor algorithm. *Int J Trends Eng Technol*, *4*, pp.65-68.
59. Madadi, P., 2009. Text Categorization Based on Apriori Algorithm's Frequent Itemsets.