From Sustainability Claims to Certified Sustainability: A Descriptive Data Analytics Approach

University of Turku
Department of Computing
Master's thesis
Computer Science
April 2026
Miisa Andersin

UNIVERSITY OF TURKU
Department of Computing

Miisa Andersin: From Sustainability Claims to Certified Sustainability: A Descriptive Data Analytics Approach

Master's thesis, 68 p.
Computer Science
April 2026

Sustainability and environmental impact are issues of extreme importance, that should affect consumer behaviour and purchase decisions. Information considering sustainability and environmental impact of a product is limited and dependent on the willingness of brands to display said information, apart from eco-labels required by legislation. A major problem in assessing a product's sustainability is greenwashing: a misleading and dishonest form of green marketing, that has several negative effects on the efforts towards a more sustainable market economy.
Inspired by the EU directive 2024/825 for empowering consumers for the green transition through better protection against unfair practices and through better information, this thesis aims to apply descriptive data analytics to real-life data in order to describe the relationship between sustainability claims in product descriptions and the real sustainability of a product depicted by the product having at least one certification, and to see if sustainability related terms are used with unsustainable products in order to mislead the customer. The research question is: "How strongly is a sustainability claim of a product description associated with the product's certification status?".
To research the relationship between generic sustainability claims and product certifications, logistic regression models were made using GreenDB data. Due to extreme sparsity of the model features in the data, meaning the lack of generic environmental claims in the product descriptions, the logistic regression models were not able to provide any reliable results and detecting greenwashing from the false positives was not possible. As previous research shows, greenwashing is very hard to identify and assess without substantial proofs of sustainability or a labeled dataset. However, statistical tests with odds ratios did show a statistically significant relationship between at least one generic sustainability claim in the product description and the product having at least one certification, meaning that a sustainability claim in the product description increases the odds of the product having a certification by 64 % in the dataset used.

Keywords: sustainability, data analytics, greenwashing

Contents

List of Figures

List of Tables

1 Introduction

Sustainability is a current theme that requires attention on a nationwide as well as on an individual level. The consequences of our actions on the environment exceed it's capability to recover, leading to grave consequences like the loss of biodiversity and climate change. Individual concerns for environmental impact lead to consumer decisions based on or taking into account the sustainability of a product, which in turn leads to the companies' interest in sustainability claims in product marketing, even when the claims are unjustified. As Persakis et al. (2025) state, the growing academic interest in greenwashing has intensified, fueled by both public demand for environmental accountability and corporate strategies that emphasize sustainable branding [1].

Unjustified sustainability claims in a product's marketing is referred to as greenwashing. Greenwashing is a problem that concerns consumers willing to consider the sustainability aspects of their purchases. The motivation for this thesis comes from the personal will to purchase products that are a better choice for the environment, while identifying and avoiding greenwashing. The EU directive 2024/825 for empowering consumers for the green transition through better protection against unfair practices and through better information states: "In order for consumers to be empowered to take better-informed decisions and thus stimulate the demand for, and the supply of, more sustainable goods, they should not be misled about a product's environmental or social characteristics or circularity aspects, such as durability,

reparability or recyclability, through the overall presentation of a product." [2] In practice it is however complicated to know if the information concerning the product is misleading or not, without a deep dive into the ESG reports of the company, and even then the individual products can be hard to assess. As the directive states, "Comparing products based on their environmental or social characteristics or circularity aspects, such as durability, reparability or recyclability, is an increasingly common marketing technique that could mislead consumers, who are not always able to assess the reliability of that information.". [2]

The sustainability claims contained in a product description can be justified by the company's actions and choices in material and manufacturing that target minimal environmental impact. Information about these choices and actions can be hard to obtain, especially for an individual consumer. The easiest way to get information to the consumer is for the brand to have their product evaluated by a third-party organization and rewarded with a certification. However a third-party evaluation can be a costly process and not all sustainable products have a certification. That is why this thesis aims to provide descriptive information on the relationship between the sustainability related terms used in product descriptions and the real sustainability of a product, in order to better understand the risk of greenwashing.

The goal of the thesis is to research the relationship between sustainability claims and real sustainability, and to provide information on the effects of the sustainability terms used in the product description on the product's sustainability status and the possible misleading use of these terms. This is done by assessing the dependency of the the product's sustainability, referring to the product having at least one certification, on the product description's generic sustainability claims by the means of descriptive data analytics. By analyzing the false positives, deeper insight should be gained into mismatch between sustainability claims and the lack of certified

sustainability status of the product, indicating possible greenwashing.

Clothing industry was selected as the topic as it is one of the most polluting industries, makes up a significant part of consumers' purchases per year and has already been the target of sustainability research with a great number of research, unlike some other industries producing everyday items for consumers.

Evaluating the sustainability of a brand or product can be a very challenging task. As Moodaley and Telukdarie (2023) say, technological improvements in artificial intelligence have presented the means to rapidly and accurately analyze large volumes of text-based information. This allows the efficient analysis of large non-heterogenous sources in order to detect greenwashing. [3] The large quantity of webshops with large quantities of products offer a lot of data for research purposes, sufficient for data analytics.

The research question of this thesis is: "How strongly is a sustainability claim of a product description associated with the product's certification status?". The aim is to describe the relationship between the generic sustainability claims and the verified sustainability of the product in the form of a certification and to gain information of possible greenwashing as a byproduct of the question. If the product decriptions do not contain any greenwashing, the environmental claims should occur only in the descriptions of products that can prove their sustainability with a certification, as is the purpose of the directive (EU) 2024/825. If the marketing of any of the uncertified products uses greenwashing, the results should differ from the frequencies of certifications and green claims, and the model should not be able to reliably predict the certification status. Instead it should produce false positives that provide information on the type of terms used in greenwashing. This is due to the fact that the goal of greenwashing is to mislead the consumer and to pass a product as sustainable without any real proof.

2 Sustainability and Greenwashing

Sustainability is an ambiguous term that encompasses a lot of themes from environmental impact to economical and social aspects. This thesis refers to sustainability as a way to ensure little to no environmental damage by a consumer product. The terms referring to these characteristics in a product's marketing are called sustainability, environmental, or green claims, depending on their content. Green or environmental claims are a specific type of sustainability marketing relating to the environmental impact, and all types of claims were used in the thesis's application to research how the claims' presence in product descriptions is associated with the product being certified. It is worth noting that the problematic claims are unjustified with lack of evidence, and generic by nature, as specified in the directive (EU) 2024/825.

2.1 Sustainability

In sustainability research, perhaps the most often used definition of sustainability is by the United Nations Brundtland Commission in 1987: "meeting the needs of the present without compromising the ability of future generations to meet their own needs". According to UN, with the increasing threat of climate change, development today must not negatively affect future generations. [4] Stöckigt et al. (2018) enlarge the UN's defitinion of sustainability by combining it with social and economic development [5].

To address and measure sustainability despite its multifaceted and even vague nature, several tools have been developed from standards to certification schemes and labels. Kesidou and Palm (2024) define sustainability standards as "voluntary governance tools that organizations use to manage their environmental, social, and ethical impacts, providing guidelines to meet criteria such as fair labour practices, supply chain transparency, resource efficiency, and environmental impact." [6] Certification schemes are criteria set by, regulated and verified by a third party organization, which can be depicted as a label that is easy for customers to identify.

Sustainability and Consumer Behaviour

Providing customers with information regarding the sustainability of the company and product is important in order to enable the customer to make conscious purchasing choices, thus enabling also a change in the market pressure to provide more sustainable choices and sift the whole mentality of purchase behaviour into a more sustainable one.

More information on the product's sustainability does not directly affect consumer behaviour in the same way, but depends on the consumer group. Stöckigt et al. (2018) showed that provided with information about the product's sustainability, consumers seem to base their decisions on sustainability-related attributes to an about equal extent as on the price [5]. O'Rourke and Ringer (2016) found a statistically significant relationship with more positive sustainability information associated with greater purchase intent. However based on their research on sustainability information on consumer behaviour in websites, many consumers are unaffected by sustainability information, and to some a more "sustainable" product may actually decrease purchase intent. On the contrary, for customers with previous interest in sustainability, the information appears to become part of their purchasing process. [7]

According to Stöckigt et al. (2018), consumers' differentiation between supposed sustainable marketing claims and proper environmentally-friendly products likely depends upon their sustainability knowledge. Consumers with a greater knowledge about sustainability are supposed to evaluate sustainability claims critically, as consumers with limited knowledge are less affected by sustainability claims. Retailers' sustainable procedures can influence consumers with a positive attitude toward sustainability and environmental concern, that are more likely to make efforts to reduce environmental impact. [5]

Transforming consumer behaviour into sustainable one is important and requires changes in manufacturing as well as customers' buying patterns. As Wojnarowska et al. (2021) state, sustainable development requires transforming consumer societies into sustainable ones, because increased demand affects growth in terms of the sale of goods and services and indirectly it affects the ecosystem in a negatice way and therefore balancing economic objectives with environmental and social objectives poses a major challenge for all parties included. [8] According to Stöckigt et al. (2018), consumers would contribute to sustainable development more if suppliers helped them by providing clear sustainability information [5].

According to O'Rourke and Ringer (2016), simply providing customers with more or better information on sustainability issues will likely have limited impact on changing mainstream consumer behavior unless it is designed to connect into existing decision-making processes. Yet information provision like public service announcements and education campaigns remains a primary strategy of governments, international agencies, academics, and non-governmental organizations. As providing more scientific information on sustainability appears to have limited impact on changing consumers' behaviour, policy makers might consider a two-stage strategy, with efforts first directed at raising consumers' awareness of issues, and then presenting these same consumers with product-level sustainability information. [7]

According to O'Rourke and Ringer (2016), sustainability measures can resonate differently depending on product type, for example deodorants are part of an "utilitarian" product group where efficacy is more important than sustainability. Decisions around these products are very resistant to influence by sustainability information. More discretionary products and products whose differences in quality is difficult to determine, are much better targets for informational campaigns to shift behaviour, as well as products that are consumed publicly. The study results indicate that consumers are focused primarily on other issues than sustainability, such as price and quality. [7]

A kind of an opposite for sustainability is materialism. Stöckigt et al. (2018) discuss materialism in their article, listing results that show materialism's negative effects on sustainable attitudes, environmental beliefs and behaviours, and is one of the root causes of environmental problems. In their study's decision-making scenarios, the environmental impact and working conditions were least important to materialists, meaning that people who decide responsibly toward the environment and working conditions are not materialistic. Results also suggest that if each product in an online shop had a simple representation of its environmental impact and of the working conditions right next to the product's price, consumers would consider this information in their purchasing decision to a similar extent as the price. [5]

According to Stöckigt et al. (2018), besides materialism, delay discounting is also negatively linked to sustainable decision making. Delay discounting describes how persons can wait patiently for a larger later reward. Regarding sustainability, the contribution to a more sustainable production is the delayed reward. [5]

2.1.1 Eco-labels and Certifications as a Measure of Sustainability

A company's sustainability can be very hard to assess. With enough information and access to valid sources, a company's environmental impact can be calculated and therefore the sustainability can be assessed. However with limited access to information, like in a customer purchase setting with only the product information available to the customer, an assessment of brand of product sustainability can be impossible. To have an objective assessment, it should be carried out by a third party and not the company itself, and delivered to the customer directly. This is the utility of third party eco-labels or certifications, that assess the company or a product, and inform the customer with their own credibility at stake.

Wojnarowska, et al. (2021) define eco-labelling as a system that informs a consumer of the environmental impact of products throughout their life cycle, that have a pre-set graphic form and constitute proof of compliance with specific norms on the part of a producer. They argue that the goal of eco-labelling is to generate demand for more desirable goods in environmental terms, and consequently to make manufacturers supply goods that live up to such expectations, and that it is regarded as one of the best tools for promoting organic products and influencing consumer-buying decisions. [8] Plakantonaki et al. (2023) see eco-labels as hallmarks of approval granted to products deemed to have fewer negative effects on the environment, and the primary purpose of eco-labeling as to encourage the manufacturing of environmentally friendly and sustainable products and to inform consumers to look for these labels before making purchases. [9] Kesidou and Palm (2024) define eco-labels as "voluntary self-regulation tools that indicate products (or processes) as environmentally preferable based on life-cycle considerations. Eco-labels signify that the product meets stated environmental and social criteria, thereby claiming it has less negative environmental (and/or social) impacts compared to similar products.

Eco-labels with independently verified, credible, non-misleading information about the environmental impacts of products, differentiates products in the marketplace with an aim to promote more sustainable production and consumption practices.". [6] Common factors to all these definitions show the most important parts of eco-labeling schemes: informing the customer of the environmental impact, marking the product as preferable to other non-certified products, and the aim to promote more sustainable consumer behaviour.

Sustainability labels are defined in the directive (EU) 2024/825 as: "Sustainability labels can relate to many characteristics of a product, process or business, and it is essential to ensure their transparency and credibility. Therefore, the displaying of sustainability labels which are not based on a certification scheme, or which have not been established by public authorities should be prohibited by including such practices in the list in Annex I to Directive 2005/29/EC. Before displaying a sustainability label, the trader should ensure that, according to the publicly available terms of the certification scheme, it meets minimum conditions of transparency and credibility, including the existence of objective monitoring of compliance with the requirements of the scheme. Such monitoring should be carried out by a third party whose competence and independence from both the scheme owner and the trader are ensured based on international, Union or national standards and procedures, for example by demonstrating compliance with relevant international standards, such as ISO 17065 'Conformity assessment – Requirements for bodies certifying products, processes and services' or through the mechanisms provided for in Regulation (EC) No 765/2008 of the European Parliament and of the Council (4). The displaying of sustainability labels remains possible without a certification scheme when such labels are established by a public authority, or where additional forms of expression and presentation of food are used in accordance with Article 35 of Regulation (EU) No 1169/2011 of the European Parliament and of the Council (5)." [2]

De Boer lists reasons for companies to use sustainability labels: they can help the company to improve their competitive position in the market and its wider environment, they can be the result of societal pressure and depend on interest in government agencies, shareholders, customers, business associations and other organizations. All these factors have the ability to raise and maintain pressure that affect the companies' profitability, which means turning environmental or social issues into economic ones. [10]

Eco-labels as a Source of Information

According to de Boer (2003), instead of sustainability labels being just messages about a product or a service, they are claims that announce particular properties or features of the product. He sees the instrument of labelling as a claim, because it refers to certain characteristics of the procedure under which the label is awarded. [10]. With growing consumer interest in sustainability, the importance of green marketing is becoming more prevalent and arguably the sentiment towards unjustified green claims is growing ever more strict as legislation and instructions for public awareness multiply.

According to Wojnarowska et al. (2021), eco-labelling plays an important role in accomplishing both sustainable production and sustainable consumption. This is due to the fact that eco-labels are expected to promote organic products and give companies a competitive advantage and to reduce uncertainty in customers and helping them choose a sustainable option. [8]

Wojnarowska et al. see eco-labelling as a system that educates consumers on the impact of products upon the natural environment throughout their entire life cycle and that provides producers with the opportunity to inform consumers about the advantages of their products [8]. Sustainability labels reveal differences between more sustainable and less sustainable practices to consumers, informing them in their

purchase decisions [10]. Informed by eco-labels, consumers can mitigate their own impact upon the natural environment and make a difference through their purchasing decisions, creating new opportunities for companies offering organic products [8]. The company's choice to adopt an eco-label is usually voluntary and as such a strategic decision to inform consumers about the environmental attributes of its products, aiming to increase sales and boost profits [11].

Consumers play an important role in decreasing the environmental damage caused by production chains by choosing "clean and green" goods and services. Thus the efficiency of eco-labels is dependent on consumers' sensitivity toward the environment (e.g. their carbon footprint). [11] According to de Boer, even though customers are interested in the sustainability factors, consumers' purchases do not straightforwardly reflect their preferences as they deal with mixed motives. To create more value for clients seeking sustainable choices, instead of focusing on the environmental or moral point, both the design and the marketing of a product should be addressed to all of the product attributes that consumers consider relevant, together with sustainability features. [10]

Yokessa and Marette (2019) list multiple examples of eco-labels transforming customer purchase behaviour towards more sustainable products, but point out that it is dangerous to generalize the market shares of products with eco-lables as they depend on many idiosyncratic parameters such as consumers' awareness, generic advertising, supply chain organization, and the accuracy of the certification process. If customers are not interested in sustainability, eco-labels effects are limited, especially as the price of a product is the main factor for limiting their purchase regardless of eco-label. [11]

Limitations of Eco-labels

Yokessa and Marette (2019) discuss the potential limitations that can damage the credibility of eco-labels. Consumers' trust is important as the sustainability characteristics are not directly visible and need a third-party verification. The credibility of eco-labels ultimately depends on consumers having faith in the third-party certifying agencies being truthful in their examinations of firms adopting eco-labels. Eco-labels should also be able to convey simple messages without betraying the complexity of environmental assessments in order to not risk information overload and incomprehension by the customer. [11]

One of the downsides of the current state and use of eco-labels is that they are becoming ever more numerous. Yokessa and Marette (2019) address the growing number and diversity of eco-labels, rooting from the growing market for green products, which in some cases limits the favorable effects that environmental information has on consumer behaviour and even producing undesirable effects. [11]

When eco-labels have a large enough market share, have credibility in the eyes of consumers, and are easy to identify by customers, they act as an easy way to convey superior state in regard of environmental impact to other uncertified products. However as eco-labels are growing in number, they can become less familiar to customers and thus instead of acting as a sign of trust they might even evoke suspicion of greenwashing.

According to Yokessa and Marette (2019), the proliferation of eco-labels is a risk as it leads to blurred signals. A multitude of labels with similar criteria for signaling sustainability makes it difficult for consumers to distinguish the most efficient ones. The weakening of the value of the greenest eco-label in the minds of consumers is a result of the profusion of eco-labels. Companies with eco-labels related to low-quality benefit from this confusion, thus making them more tempting for the companies. [11]

In their article "The bunch of sustainability labels – Do consumers differentiate?", Janßen and Langden researched consumers' perception of different sustainable attributes labelled on milk and found three different customer profiles regarding preferences on sustainability labels in groceries. According to Janßen and Langen, no general statement can be made regarding the question whether different sustainable aspects complement or substitute each other. However their study clearly shows that dispite the three different customer segments, almost 85% of the market could be satisfied with a universal sustainability label, instead of the customers preferring more labels on a product. [12]

The credibility of eco-labels for signaling high quality ultimately depends on whether consumers have faith in the third-party certifying agencies truthfully reporting their examination of firms adopting eco-labels. However, public and private certifying agencies cannot always perfectly determine quality at a reasonable cost or truthfully report the environmental impact of products. [11]

According to Yokessa and Marette (2019) eco-labels should be considered to complement standards banning/limiting non-green products and taxes/subsidies on green products. This combination could overcome the limitations that eco-labels currently have. [11] Kesidou and Palm (2024) list four key dimensions to reduce consumer uncertainty and enhance eco-label effectiveness: standardisation of environmental attributes such as carbon footprint and biodiversity, improved transparency in monitoring and certification processes by third-party verification, lifecycle and value chain considerations, and effective communication of eco-label information. They also list four recommendations to strenghten the effectiveness of eco-labels in the fashion textile industry: ensuring consistency in environmental standards, adopting a more systems-based approach, strenghtening eco-label standards beyond third-party verification, and improving communication strategies. [6]

Certifications can increase the perceived value of eco-friendly brands and con-

sumer willingness to pay. According to the theory of eco-opportunism, this can lead to free riding and greenwashing, where products are falsely advertised as sustainable but fail to meet certified standards. [13]

Different Types of Eco-labels

There are different categories of eco-labels that vary depending on various factors, like product or target group, type of evaluator, degree of obligatoriness, and strictness of criteria. Ideally, when using eco-labels as a tool of assessment, each eco-label should be categorized and assessed. For example Yokessa and Marette (2019) characterize eco-labels using various criteria such as communication channel (e.g. business-to-consumer, business-to-government), category of goods and services targeted, environmental attribute (e.g. sustainability, biodiversity), ownership (e.g. private, public, non-profit), mode of governance (e.g. voluntary vs. mandatory), and scope (e.g. regional, international) [11].

Eco-labels can be sponsored by governments, non-governmental organisations, or business associations, each offering different levels of credibility. Credibility depends also on monitoring: some are internally monitored through self-assessment, others use second (sponsor) or third-party (an independent party) verification. [6]

Yokessa and Marette (2019) also discuss mandatory eco-labeling, saying that firms have to incentive to voluntarily eco-label their goods that have negative credence attributes, which is a major motivation for justifying a mandatory eco-label. They see eco-labels being the most beneficial when combined with other regulatory mechanisms: policy instruments like standards banning/limiting non-green products, taxes on non-green products and/or subsidies on green products. [11]

De Boer (2003) claims that policymakers support or regulate labelling schemes in order to address the economic interest of consumers by prevention of misleading advertising or deceptive environmental claims, and to achieve sustainability objec-

tives by government policies, particularly by promoting the design and marketing of environmentally sound products or services. Labelling and certification schemes are tools for policymakers for creating incencitives for businesses to change the market in a more sustainable direction. Governments can establish mandatory labelling laws, regulate claims through legal definitions of specific terms, provide services to support voluntary labelling, and link the terms of purchase to labelling and certification schemes. Another group that can use supporting or criticizing labelling schemes as a tool towards sustainability are environmental or social NGOs. [10]

The Quality of Eco-labels

The quality of eco-labels differ, depending on the strictness of criteria and the strictness of evaluation. De Boer (2003) divides sustainability labels into two strategic categories: labels as a benchmark to achieve ideals and labels as a bottom line to avoid ills. For example the EU eco-label is an example of the environmental labels that, especially the multi-issue eco-labels, are often designed as a benchmark of excellence. Social labels he sees as often designed as becoming the bottom line of the market. De Boer points out that because of these differences in strategy, sustainability labels cannot simply replace the existing environmental and social labels. [10]

Due to the nature of eco-labels existing in relation to the customer, providing information and targeting to convince the customer of the product's superiority, measuring the effectiveness of eco-labels is not simple. According to de Boer (2003), it is not feasible to draw generalizing conclusions on the effectiveness of labelling and certification schemes, as they are closely connected with the pressure generated by all kinds of actors in society to change the consumption patterns in a more sustainable direction, which in turn is not the same in all sectors and industries. [10]

2.1.2 Sustainability in Clothing Industry

As well as one of the oldest and most diverse manufacturing industries, the textile industry is one of the most polluting [9]. Sustainability in the clothing industry has been a subject of discussion and scrutiny for a long time, as the industry is one of the biggest consumer markets and has clear consequences on the environment and the health of the industry workers.

Scrutiny surrounding the fashion textile industry dates back to the dawn of the Industrial Revolution, and the environmental impact of the industry began to attract attention already from the 1850s with social concerns soon following [6]. Due to a longer history of sustainability concerns, there are more research and sources directed directly to consumers publicly available concerning sustainability in the clothing industry, than for example concerning the cosmetic industry. Research on consumer behaviour in the textile fashion industry, especially concerning sustainability related beliefs and attitudes has increased substantially in the past decade, indicating consumers becoming more conscious of the impacts of their consumption choices and prompting a growing demand for information to support ethical decision making [6].

Plakantonaki et al. (2023) claim that the textile and fashion industries bear a negative impact on the environment and contribute significantly to water, air, and solid waste pollution, especially as clothing purchases have increased dramatically over the last decades. The textile manufacturing process involves the use of numerous chemicals that are harmful to both humans and the planet. In addition to generating a large amount of toxic waste and greenhouse gasses, textiles have been identified as unsustainable products due to their entire life cycle. The textile industry is a major polluter, and textiles are characterized as unsustainable products from raw material cultivation to manufacturing, producing solid waste, pollution and requiring a large amount of water. [9] The negative impact of the textile and fashion

industry upon the environment is exerted during both production and consumption phases [6]. A more sustainable textile sector requires major brands implementing sustainable manufacturing practices, eco-friendly textiles, restricted substances, and eco-labeling. [9]

According to Kesidou and Palm (2024), fashion textile industry faces unique challenges in implementing sustainable practices, illustrated by the ongoing tension between ethical awareness and the demand for low-cost, fast-moving fashion products. The sustainability ethos promoted by eco-labels is often opposed by the very nature of fast fashion. [6] The fashion and textile industry is a sector known as one of the largest industrial polluters worldwide, and greenwashing is a prevalent issue inside it. The multi-trillion-dollar industry's mass consumption stems from trend-driven, low-cost production models and the availability of numerous distribution options impacting consumer behaviour. [14] Sustainability issues are prevalent throughout the entire fashion supply chain, including intensive use of chemicals, water, energy, and undegradable fabrics which leads to problems such as the rapid increase in damaging plastic microfibres, hazardous ingredients pollution, and large amounts of waste [15].

According to Kesidou and Palm (2024), for the industry to fully embrace ecolabels and other sustainable practices, a shift in consumer attitudes and expectations is required in addition to industry operations [6]. According to the legislations by governments promoting ecolabels and sustainability standards, textile manufacturers and product designers must pay special attention to sustainability standards, making the product ecologically responsible at all stages of its lifecycle and taking into account ecological, social and economic concerns [9].

Eco-labels in the Clothing Industry

Kesidou and Palm (2024) discuss eco-credentials and sustainability in fashion and textile industry. Eco-credentials are standards that help firms enhance their environmental performance. They encompass standards such as ISO standards for standardising firms' internal procedures and ecolabels for products in the marketplace. [6]

According to Kesidou and Palm (2024), despite actions of the textile and fashion industry towards sustainability, its negative environmental impacts are increasing. According to them, eco-labels play a critical role in shaping sustainability practices within the industry, by setting benchmarks for environmental attributes, promoting transparency, and fostering lifecycle and value chain consideration. Eco-credentials as a form of voluntary measures have proliferated in recent years as part of the industry's response to concerns about environmental impact. [6]

Eco-labels in fashion textile industry have potential to help align environmental priorities and guide consumer purchasing decisions. They serve as a signal of a product's environmental credentials, guiding consumers toward more eco-friendly choices. Their strenghts are e.g. standardization, and enhanced transparency and lifecycle integration. Rigorous third-party verified certification processes contribute to transparency and trust in sustainability standards. By covering various lifecycle stages, eco-labels provide a comprehensive view of environmental impacts. [6].

Almeida (2015) sees eco-labels in textile products as a way to show the consumer that the products are safe in terms of human health, produced with environmentally friendly materials and technologies corresponding to real ecology, produced with regard to the health and safety of the workers, and produced with respect to social criteria in terms of the human rights of workers. Different eco-label systems impose criteria regarding certain performance quality levels, affecting the entire textile chain. Eco-labeled products should respect restrictions beyond the legislation in

more developed countries, independent of the country where textiles are produced. [16]

Proliferation and diversity of eco-labels is a problem also in the clothing industry. According to Plakantonaki et al. (2023), there is no systematic techniques for determining which ecolabeling scheme is best for a certain clothing product, and there is no stantardized method for a textile manufacturing company to choose an appropriate eco-label, as different eco-labels have different impact criteria depending on their region. [9]

Kesidou and Palm (2024) dissect the problems of eco-labels in the clothing industry in more detail: eco-labels' limitations in fashion textile industry are inconsistent environmental attribute selection, lack of scientific integration, and vague criteria that hinder effective cross-label comparisons and limit consumer comprehension. Consumer distrust is increased by transparency gaps in certification processes, particularly regarding certifier ownership structures. The proliferation of eco-labels obscures the process for firms trying to select the most suitable eco-credential. Eco-labels focus on the product-specific environmental impact, but not all cover the entire value chain. The variability in scope affects their suitability for assessing the overall environmental footprint. Other weaknesses of eco-labels are lack of consistency, fragmented sustainability approaches, and communication issues. Vague language and insufficient transparency undermine the educational value of ecolabels and leads to consumer distrust. [6]

2.1.3 Greenwashing in the Clothing Industry

As any other consumer product manufacturing industry, the clothing industry is guilty of greenwashing. The nature of fast fashion as an extremely polluting and unsustainable business, combined with growing demand for sustainable products and consumer eco-consciousness creates a temptation for companies to perform green-

washing. According to Lu et al. (2022), it is undeniable that many fast fashion companies tend to cover up the unsustainable part of their business activities through fake green marketing campaigns to gain more potential consumers [17].

According to the Changing Markets Foundation, the fashion industry is one of the least regulated sectors in the world. The certification schemes and the inherent lack of accountability within them are a key part of the greenwashing machinery inside the fashion industry. Certification schemes do exist partially as a genuine attempt to move towards sustainability in the absence of environmental legislation, but they also enable the proliferation of greenwashing on a remarkable scale, as certification labels on individual products assure customers that they can shop guilty free, and brands can proudly communicate their membership of various voluntary initiatives. The level of influence of fashion brands in the cerfitication initiatives and the lack of independent oversight means that they end up promoting industry interests. [18]

2.1.4 Green Marketing

"Sustainability" and "green" are terms used to depict a lesser environmental impact, often used interchangeably, but they do contain a difference of meaning. As Alkhatib et al. (2023) point out, green marketing and sustainability marketing are not the same. While green marketing focuses on promoting environmentally friendly products, sustainability marketing takes a broader perspective involving the inclusion of the entire community, including its social objectives and efforts towards environmental preservation. [19] In that regard, green marketing is a part of sustainability marketing, but with a more limited focus. Greenwashing on the other hand is a misleading extension of green marketing, encompassed in the theme of sustainability marketing.

The European Council defines green or environmental claims as explicit written or oral statements by businesses, aimed at consumers, which imply that their

product, services, or the organisation itself has a neutral, reduced or positive environmental impact, or no environmental impact at all. Such claims give consumers the impression that their purchase contributes to a more sustainable economy [20]. Alkhatib et al. (2023) define green marketing as minimizing a product's environmental impact through product redesign, sustainable manufacturing, and integrated marketing campaings, aiming to promote eco-friendly products and meet the demand for sustainable consumption as well as to position environmentally friendly products in the market and appeal to environmentally conscious consumers [19].

According to Ni (2024), the core of green marketing lies in balancing economic benefits with environmental responsibility, optimizing resource utilization, and reducing pollution to achieve long-term competitive advantage [21]. According to Badhwar et al. (2024), green marketing encompasses the process of lessening the environmental footprint of the products through redesigning, sustainable production, and well-coordinated marketing strategies [14].

For Moravcikova et al. (2017) the main goal of green marketing is to present consumers with the importance of protecting the environment in the context of product consumption, placing an emphasis on building long-term relationships based on both sides of communication, not only with customers but also with other stakeholders and creating the natural need to be environmentally responsible [22]. Succesfull green marketing while avoiding the stigma of greenwashing requires companies to take substantial sustainability measures and adopt a factual approach without exaggeration or concealment [17].

Green marketing is a growing field in marketing following the rise of consumer interest in sustainability. According to a report "Environmental claims in the EU" published by the European Union, 80 % of shop pages and advertisements of products/services contained at least one implicit or explicit environmental claim. 53 % of the environmental claims were potentially misleading. 44 % of clothing contained

environmental claims. 176 unique logos were identified in the study. [23]

2.2 Greenwashing

Greenwashing is a part of green marketing, where the subject is environmental impact and sustainability but the actions are defined by mislead and dishonesty. It is a phenomenon that intertwines several aspects of companies, consumers, financial and market aspects, environmental impact, and society. Due to its multifaceted nature, the definitions vary depending on the point of view and emphasis on a selected aspect.

Nemes et al. (2022) define greenwashing as the practice of falsely promoting an organisation's environmental efforts or spending more resources to promote the organisation as green, than are spent to engage in environmentally sound practices. Thus, greenwashing is the dissemination of false or deceptive information regarding an organisation's environmental strategies, goals, motivations, and actions. [24] Alkhatib et al. (2023) define greenwashing as a type of marketing spin in which the concept of green marketing is used to deceive the public into believing that a company's goods, goals, and policies are environmentally beneficial when they are not. [19] Delmas and Burbano (2011) define greenwashing as the intersection of two firm behaviours: poor environmental performance and positive communication about environmental performance. They divide it into firm-level greenwashing, where companies mislead consumers regarding their environmental practices, and product-level greenwashing, where companies mislead consumers about the environmental benefits of a product or service. [25]

Addressing the problem of ambiguity around different forms of greenwashing and creating a tool for assessment and identification, Nemes et al. (2022) present an integrated framework to be used by actors of any kind to avoid greenwashing, including corporations, governments, and other organisations. It provides a structured way

to ask questions about the different varieties of greenwashing to evaluate whether the organisation under assessment could be considered as engaging in greenwashing or not. The framework collects in one set of greenwashing themes a broader range of indicators linked to possible sources of greenwashing and could support effective sustainability policies and genuine green marketing and communication strategies at the level of any organisation. [24]

2.2.1 Types of Greenwashing

Greenwashing can be hard to identify as it can appear in many different ways: in the form of product design, utilizing stereotypes of a sustainable product in product or packaging design, using vague or misleading language in product descriptions and in association with symbols and certifications that give credibility or positive image on the product. Even ecolabels can be taken advantage of in greenwashing. Companies can utilize a selfmade symbol that gives the impression of a certificate, or they can have a less strict certificate scheme that gives credibility without too much effort in environmental change, and could be combined with enhanced claims that are not factual for an even stronger effect. According to Yokessa and Marette (2019), consumers may be more influenced by the marketing around a label than the environmental quality it represents [11], especially with the profusion of eco-labels and in the case of lesser known eco-labels.

Popular in the field of sustainability and greenwashing, the marketing and consulting organisation TerraChoice's Seven Sins of Greenwashing determines guidelines for identifying companies' greenwashing behaviour:

1. Hidden trade-off: suggesting a product is "green" based on an unreasonably narrow set of attributes without attention to other important environmental issues.

2. No proof: an environmental claim that cannot be substantiated by easily ac-

cessible supporting information or by a reliable third-party certification.

3. Vagueness: so poorly defained or broad claim that its real meaning is likely to be misunderstood by the consumer.

4. Irrelevance: making an environmental claim that may be truthful but is unimportant or unhelpful.

5. Lesser of two evils: claims that may be true within the product category, the product category having greater environmental impacts as a whole.

6. Fibbing: making environmental claims that are simply false.

7. Worshiping false labels: a product giving the impression of third-party endorsement by words or images, where no such endorsement actually exists.

[26]

According to Alizadeh et al. (2024), the most common greenwashing-like behaviour among fashion brands are "misleading", "concealing", and "vagueness", with "overselling" or exaggeration and "irrelevance" following. [27]

In addition to direct greenwashing, Badhwar et al. (2024) depict indirect greenwashing that takes place in the company's relation with its suppliers. Indirect greenwashing occurs when a company's association with suppliers or partners who claim to be sustainable is misleading, where suppliers claim to be environmentally conscious, but their practices might not align with their claims. Another type is vicarious greenwashing, where a company appears environmentally responsible in its marketing, but this image is a result of its association with suppliers that lack genuine sustainability efforts. [14]

Greenwashing is a real problem with direct consequences, as it undermines the real efforts for sustainable products, and creates unfair competition between companies putting effort in real sustainability and companies looking for profit without any

regard for their environmental impact. In addition to direct consequences, greenwashing has undirect consequences in the form of creating vagueness and mistrust in the consumer market.

According to Badhwar et al. (2024), businesses use vague terms, keywords (e.g. eco-friendly, chemical-free, organic, and sustainable), and tactics to spread sham marketing messages. This in turn drives misconceptions of green-related terminology and fills the gap between customer expectations and information that businesses disclose. Deceptives practices, ambiguous language, and lack of transparency continue to mislead consumers and emphasize the critical need for increased accountability and a reevaluation of current sustainability standards in the fashion sector, despite the existence of certifications and initiatives. [14]

2.2.2 Drivers of Greenwashing

At the same time as the amount of green products, services, and firms has expanded, the amount of firms engaging in greenwashing has increased [25]. Companies are tempted to use greenwashing, as it does not require much effort, the financial benefits can be significant, and because getting caught is a relatively small risk, as the proof of greenwashing can be hard to get and the surveillance is not automatic. Financial benefits include attracting more customers and thus market leverage, achieving the benefits of a positive environmental image with less money, and saving the money needed to meet regulation demands etc.

According to Persakis et al. (2025), in previous research greenwashing is viewed as a strategic move used by companies to gain market leverage by manipulating environmental information. Greenwashing reflects a strategic approach where companies distort environmental achievements to attract eco-conscious consumers, gain market share, or justify prerium pricing without meaningful operational changes. [1] According to Moravcikova et al. (2017), accepting the principles of green marketing

increases the value of the company's products, the company gains a competitive edge, improves its image, gets to new markets and is prepared to cope with the environmental pressures of stakeholders [22].

In their study on drivers of greenwashing, Delmas and Burbano (2011) focus on why firms with poor environmental performance want to communicate positively about their environmental performance. These drivers they organize into three levels: external, organizational, and individual. External drivers include regulators and non-governmental organizations that perform surveillance, and market actors like consumers and investors. Individual level contains psychological factors affecting decision making, like optimistic bias and narrow decision framing. Organizational level contains the factors inside the company that affect the company culture and decision making, like organizational inertia, firm characteristics and effectiveness of intra-firm communication. [25]

The regulatory context is a critical driver of greenwashing due to the limited punitive consequences of greenwashing. At the time of the article being written, current regulatory environment was the key driver of greenwashing as regulation of greenwashing was extremely limited in the US and enforcement of such regulation was highly uncertain. For multinational corporations, variation in regulation across countries and complexity regarding appropriate jurisdiction of cross-country practices contributes to a particularly uncertain greenwashing regulatory environment. The regulatory context is also an indirect driver of greenwashing as it influences the other drivers of greenwashing with its effect on information. The information about firm greenwashing and environmental practices influences the consumer and investor demand for green products, services, and firms, while the lax and uncertain regulatory context significantly contributes to the availability and reliability of this information. [25] Because the regulatory context is such an important factor in greenwashing, regulation on the level of the European Union in the form of a special

directive is deemed necessary.

Delmas and Burbano's (2011) classification's market external drivers contain consumer demand, investor demand and competitive pressure [25]. Consumer demand refers to environmentally consciours clients looking to make environmentally conscious purchases. According to Fella and Bausa (2024), consumers have a limited ability to identify greenwashed products and honest green products. When consumers mistake a greenwashed product as an honest green product, companies pretending to be green may benefit of the confusion [28]. As consumers are more willing to pay a prerium price for eco-friendly goods and have an intention to purchase green products at a higher price compared to non-green products [19], eco-conscious consumers are a tempting target group for companies willing to make more profit with misleading marketing. Having a green image and offering high-quality green products contribute significantly to customer satisfaction, which also indicates consumers' willingness to pay higher prices for environmentally friendly products of superior quality [19].

Likewise to Delmas and Burbano (2011), Yang et al. (2020) identify governmental policies, competitive pressure, and market opportunities as the drivers of greenwashing. The market pressure or competitive pressure refers also to other companies in the market, that affect the behaviour of a firm via certain channels through actions undertaken by other firms to reach the same group of consumers in the market. The growing environmental awareness of consumers is seen in Yang's classification as a market opportunity, which is identified by a new demand that a firm can meet, as it is not supplied by competitors. [29]

Another reason for greenwashing stems from a passivity in environmental communication or pre-emptive control of negative brand image. As Delmas and Burbano (2011) say, firms with poor environmental performance remain either silent or try to represent their bad environmental performance in a positive light, as it would be

counterproductive for a firm to actively communicate negatively about its bad environmental performance, thus their communication ranges from no communication to positive communication [25].

2.2.3 Effects of Greenwashing

Greenwashing has various widely recognized negative effects from unfair competition between brands to undermining consumer trust and deteriorating real efforts to hinder the environmental degradation. However as the phenomenon of greenwashing is hard to detect with certainty, its effects are also complicated to measure.

Greenwashing erodes consumer trust and promotes skepticism, particularly when eco-claims lack transparency or consistency [1]. According to Bao et al. (2025), greenwashing can not only undermine consumer trust but also hinder genuine sustainability efforts, as consumers realising that they have been misled by green claims may become sceptical of all green marketing, even from genuinely sustainable companies, reducing the overall effectiveness of green marketing campaigns [30]. The deterioration of trust stemming from greenwashing is not only limited to consumer trust. The negative consequences of greenwashing touch various actors, including consumers, companies, stakeholders, the environment, and society at large by misleading consumers and increasing mistrust, and harming corporate financial performance by undermining corporate reputation [27].

Through consumer trust deterioration, financial effects concern all companies regardless of greenwashing behaviour. According to Lu et al. (2022), consumers' perception of greenwashing in the fast fashion industry has a direct negative effect on their green purchase intention: the more consumers perceive a company's greenwashing behavior, the more it will weaken their willingness to purchase related products [17]. When consumers think of an honest green product to be greenwashed, they may unintentionally penalize genuine companies that actually try to improve

their sustainability performance. [28]

According to Delmas and Burbano (2011), not only can greenwashing erode consumer confidence in green products, but also negatively affect investor confidence in environmentally friendly firms. As confidence erodes, stakeholders become reluctant to reward companies for environmentally friendly performance. [25] The consequences become extreme in the long term, as fewer investors or stakeholders are willing to invest in production of green products for a market [29]. This, in turn, encourages firms to engage in harmful behaviours [29] and increases the incentives for firms to engage in environmentally detrimental behaviour, creating negative externalities and thus negatively affects social welfare [25] and society as a whole [29]. It also entails some risks to greenwashing firms themselves, when consumers, non-government organizations or government entities question firms' claims. [25]

The negative effects of greenwashing become clear to the perpetuating companies themselves when companies get caught of performing greenwashing. The exposure of greenwashing may damage the brand's reputation and have a series of collateral effects on the consumer market [17].

In addition to negative effects on the confidence of shareholders and consumers in green products, greenwashing is also disadvantageous for the environment and consumers' health [29]. Exaggerated claims of work in favor of sustainability and the environment on behalf of greenwashing companies create a misadvantageous and unfair competition setting for companies that are truly interested in their environmental impact, thus undermining their possibilities of investments in sustainability.

2.2.4 Directive (EU) 2024/825 of the European Parliament: Empowering Consumers for the Green Transition through Better Protection Against Unfair Practices and through Better Information

As greenwashing has become a serious matter that hinders the efforts against climate change and environment degradation on a global scale, the European Parliament has responded by enlargening the already existing regulation againts misleading marketing with a special directive targeting greenwashing. As a continuation for previous legislation, the goal of the directive (EU) 2024/825 is to put an end to greenwashing, promote environmentally-friendly decisions, create a circular economy that reuses and recycles materials, and to provide more information to consumers on the durability of products they buy in order to better protect consumers' rights [31].

Besides EU, greenwashing has also provoked government invervention on a national scale. For example in France, the Ecological Transition Agency has published an anti-greenwashing guide designed to help companies and individuals to assess their communication and environmental performance to avoid greenwashing [32]. As a part of the campaign, they have released a web poll with questions concerning communication in order to help users to evaluate their own risk of greenwashing [33].

As verifying the reliability of green claims of an increasing amount of green brands becomes complicated, the European and national authorities have provided that green claims must be true, reliable, verifiable, and comparable [34]. In practice this means that all green marketing has to be backed up or justified by actual proof, like a certification, and that generic environmental claims, or so-called "buzzwords" like "eco-friendly" or "nature's friend" are banned unless justified by aforementioned

proof. The examples of banned generic environmental claims listed in the directive are used as the labels of the logistic regression model in the application part of the thesis, where they are identified in the product descriptions.

The reason for banning generic environmental claims on products without proof is that all information on a product's impact on the environment, longevity, reparability, composition, production and usage should be backed up by verifiable sources. Together with generic environmental claims without proof, the directive also bans claims that a product has a neutral, reduced or positive impact on the environment because the producer is offsetting emissions, and sustainability labels that are not based on approved certification schemes or established by public authorities. [31]

The directive (EU) 2024/825 states that "In order to contribute to the proper functioning of the internal market, based on a high level of consumer protection and environmental protection, and to make progress in the green transition, it is essential that consumers can make informed purchasing decisions and thus contribute to more sustainable consumption patterns. That implies that traders have a responsibility to provide clear, relevant and reliable information. Therefore, specific rules should be introduced in Union consumer law to tackle unfair commercial practices that mislead consumers and prevent them from making sustainable consumption choices, such as practices associated with the early obsolescence of goods, misleading environmental claims ('greenwashing'), misleading information about the social characteristics of products or traders' businesses, or non-transparent and non-credible sustainability labels." [2]

The directive was set in March 2024, and the member states of the European Union must apply it from 27 September 2026 [2]. Hence it should not yet have an effect on the product descriptions contained in the technical research part of the thesis. Inside the EU, now is the optimal time to accumulate real data on green marketing for greenwashing research before the directive's effects start to show.

3 Machine Learning Methods for Detecting Greenwashing in Previous Research

An important part of data analytics, machine learning provides efficient methods for making predictions based on training data. A lot of research has applied machine learning tools on researching sustainability and greenwashing.

According to Kobti et al. (2021), the rise of fake news has put much attention on developing methods and tools to detect false claims. Greenwashing is a domain where false, inaccurate and misleading information plays a major role. They see automatic detection of green claims as necessary to facilitate the detection of greenwashed claims. [35]

With its recent rapid progress, machine learning has become a convenient method for research on sustainability of brands and products. It offers ways to overcome the challenges stemming from various different types, forms and styles of sources enabling assessment of sustainability and green marketing. According to Persakis et al. (2025), machine learning and AI models offer advanced methods for identifying greenwashing by analyzing textual and quantitative data across corporate disclosures and ESG metrics [1]. As Kobti et al. (2021) state, automatic green claim detection is still an underexplored problem from a computer science perspective; consequently,

a new AI model must be designed, trained, and appropriately evaluated. [35]

Previous research focuses on textual sources containing information on the company's sustainability efforts and environmental impact, often required by legislation or provided to inform the public or shareholders. Perhaps the most researched, ESG reports are reports on a company's environmental, social, and governance impacts that help stakeholders to make informed decisions and provide corporate transparency and accountability, mandatory in several countries. As Kang and Kim (2022) state, many companies produce an annual sustainability report intended for stakeholders and the public, enumerating the goals and degrees of achievement of the company regarding sustainable development. As extracting key information from the long reports can be complicated, many researchers have attempted to analyze the concepts and messages from sustainability reports using various natural language processing methods. Because of a wide disparity between reports, researchers have attempted to analyze sustainability trends, key messages, and focus areas. Various tools have been used to determine the presence of certain words, themes, or concepts within the reports. [36]

In their own research, Kang and Kim (2022) use sentence similarity method and sentiment analysis to show thematic practices and trends as well as a significant difference in the balance of positive and negative information in the reports across companies. Their pre-trained language model analyzes the contents of the sustainability reports, and through the sentence similarity method they made a quantitative measurement of the contents according to a predefined theme structure. The sentence similarity method overcomes the short-comings of word frequency-based methods, popular in previous research, that do not incorporate the context of the text. [36]

Chen and Ma (2024) used text mining technology to develop an ESG greenwashing detection tool by first creating an ESG lexicon and then searching ESG reports

for keywords and matching them with the created dictionary to achieve accurate extraction of the annual ESG-related corpus of an Enterprise using Word2vec and TF-IDF. Finally they calculate a greenwashing score by sentiment analysis. [37]

Within the European Union, legislation requires companies to provide information considering their social responsibility, sustainability, and environmental impact, which is also fertile ground for sustainability research. The Corporate Sustainability Reporting Directive (EU) 2022/2464 is an integral part of the EU's Green Deal initiative, that obligates businesses to disclose how their operational activities influence various sustainability indicators [38]. These indicators include the impact of their activities on people and the environment, and what they see as the risks and opportunities arising from social and environmental issues. Companies subject to the CSRD have to report according to European Sustainability Reporting Standards (ESRS) [39].

Li and Zhao (2021) use CSR reports to provide a holistic perspective of fashion companies' sustainable development and investigate the sustainability practices of global fashion companies by implementing Dictionary approach text classification method, combined with Latent Dirichlet Allocation, a computer-assisted topic modeling algorithm, on CSR reports to detect and summarize the themes and keywords of detailed practices disclosed in the reports. [15]

Despite legislations' obligations for sustainability reports, the form of the resulting products can vary significantly, which poses problems for the analysis of grand masses in sustainability research. Mahdavi et al. (2024) propose a clustering-based algorithm to automatically detect sustainability objectives in heterogenous reports. It extracts text elements of heteronegous sustainability reports and groups them into informative text blocks to create an extensive labeled dataset with domain expert annotations. Finally it fine-tunes a pretrained transformer model on the dataset to predict sustainability objectives in any new report. [40]

In addition to regulatory reports, other possible sources include webshop descriptions, customer reviews, and social media. Using social media as their source, Huang and Li (2024) proposed a mechanism that combines multimodal analysis with social intelligence to identify common traits among greenwashing claims from corporate sustainability claims in Twitter. Their classification model assesses the significance of image features and social interaction features like replies, and number of likes and retweets in detecting greenwashing. [41]

In their article on uncovering sustainability insights from Amazon's eco-friendly product reviews, Maarif et al. (2023) use natural language processing techniques, including sentiment analysis, key term extraction, and topic modeling to extract meaningful insight from large amounts of textual data in the form of product descriptions and consumer reviews. To understand consumer sentiments and preferences related to sustainability aspects, they integrate the natural language processing approach with correspondance analysis that shows the interplay between eco-friendly product features and consumer sentiments, revealing underlying relationships and patterns. [42]

The research of Maarif et al. (2023) on customers' product reviews and perceived sustainability of products shows dominance of positive sentiments regarding eco-friendly product attributes, indicating that a large portion of consumers appreciate sustainable products. Their findings indicate that users evaluate sustainable products not only based on their eco-friendly attributes but also based on how these attributes intersect with other product qualities. The terms related to "eco-friendly" or "sustainable" labeled on any kind of product are evidently perceived positively among consumers, indicating their acceptance of the product value that is designed with environmental fashion. Their examination of positive sentiments identified specific satisfaction aspects, such as product quality, usability, appearance, and value for money. [42]

4 Measuring the Association Between Sustainability Claims in Product Descriptions and Product Certification Status

While making a purchase decision of a clothing product in a webshop, the sustainability information available to the customer is usually limited to the product description and possible certifications. The purpose of the research done in this thesis is to apply descriptive data analytics to product data in order to gain insights into the relationship between the generic sustainability claims in product descriptions and the sustainability of the product. This provides information on the credibility of the sustainability claims and the risk of greenwashing, as a strong association can be interpreted as only sustainable products using sustainability related terms in their descriptions, and a weak association can be interpreted as unsustainable products using sustainability related terms in a misleading way.

The goal is to produce a model that can predict if a product is sustainable or not based on the sustainability related terms used in the product description. A target variable that depicts or quantifies the sustainability characteristics of the product, and a sufficient quantity of learning data that contains enough features that predict

the target value to make a signal strong enough for predicting the target values of the test data set.

Clothing was selected as the target product group as the fashion textile industry is arguably one of the most polluting industries, and makes up 4 % of total household consumption expenditure in the European Union [43]. On an industry level, a lot of attention has been brought over the years to sustainability issues, resulting in research and evaluation of brand sustainability as well as focus on environmental values in the brands' marketing strategies, and databases providing data for research.

The logistic regression models produced in this thesis aim to identify a product's certification status based on the green claims in the product description follows the path of previous research on machine learning approaches on greenwashing. The closest example is from Kobti et al. (2021), who proposed a supervised method to detect green claims made by companies in social networks. Their results show that it is possible to recognize green claims with considerable performance from short texts using a binary logistic regression classifier. [35] In this study, logistic regression is applied to product descriptions to predict the certification status of the product, as it is a simple model and relatively easy to interpret.

As previous research has shown, there are several problems in identifying greenwashing directly with the means of machine learning, and previous research has often focused on smaller tasks such as identifying green claims [44]. The same difficulties became apparent in this study, requiring gradual modifications to the research setting from more elaborate test setting towards a simpler and more limited setting as explained in chapter 4.5.

4.1 GreenDB

In order to assess the association of sustainability terms in product descriptions and the sustainability of the product with data analytics, sufficient training data on

products acquired from webshops is necessary. As no database containing labeled data of greenwashing exists [44], the only way to conduct research is to get the data from real webshops. As web scraping is often prohibited by the terms of use of web stores, GreenDB provides a way to get enough data for data analytics, preprocessed into tabular form.

GreenDB database by Jäger et al. is a product-by-product sustainability database publicly available for research purposes. It contains the product attributes and sustainability information of clothes and electronics collected from several European web stores in Great Britain, France and Germany. The transparently evaluated sustainability information enables the ranking of products based on their sustainability. [45] The benefits of GreenDB are the great quantity of information, real-life data and the diversity of information collected of each product, containing images, descriptions, prices, certifications, etc.

In this thesis the GreenDB data was limited to clothing items in order to provide a coherent set of learning data. The data was divided into three subsets by market region, where the product descriptions of each subset were written in either English, French or German, depending on the market area. A language specific subset allows for detection of generic sustainability terms in the product description, which is an essential step in the prediction of certification status and detection of possible greenwashing suspects.

4.2 Targets: Certifications

In order to research the relationship between green marketing and sustainability, the machine learning model has to predict a target value, which should depict or encompass aspects of sustainability. When searching for product information regarding the environmental aspects of a product, the customer is often left with the sustainability claims in the product description, possible certifications, and eco-labels.

According to the directive (EU) 2024/825, the use of generic environmental claims should be prohibited in marketing when recognised excellent environmental performance cannot be demonstrated. Whenever the claim is justified by providing the specification of the environmental claim clearly and in prominent terms on the same online selling interface, the claims are not considered to be generic environmental claims. [2]

The requirements given in the directive are elaborate. According to the directive, the specification of the environmental claim or recognised excellent environmental performance, that renders the claim from generic to justified, can be demonstrated by compliance with Regulation (EC) No 66/2010 or with officially recognised EN ISO 14024 ecolabelling schemes in the Member States of the European Union, or by corresponding to top environmental performance for a specific environmental characteristic in accordance with other applicable Union laws. [2] This means that to fulfill the requirements of the directive when using environmental claims, it is not enough to provide a third-party eco-label, but the label has to be under or aligned with EN ISO 14024 standard.

When accompanied by a certification scheme, the claims should correspond the criteria of evaluation in the certification scheme. The directive specifies that "a trader should not make a generic claim such as ‘conscious’, ‘sustainable’ or ‘responsible’ based exclusively on recognised excellent environmental performance, because such claims relate to other characteristics in addition to environmental characteristics, such as social characteristics.". [2] To evaluate if a claim is a generic environmental claim or justifiable green marketing, also the provided specification like a third-party eco-label should be assessed, by comparing it’s criteria and compliance with EN ISO 14024 standard in order to accept the use of the environmental claim. However the other means proposed by the directive, top environmental performance in compliance with other union legislation can be even more complicated to assess.

Certification schemes and eco-labels provide thus the easiest way to verify a product's sustainability and the assessment of the justification of environmental claims, that is also accepted by the directive when combined with enough scrutiny. This is why the certification status was chosen as the target variable of the machine learning model.

GreenDB has a list of eco-labels granted for each product. The possible values of the sustainability labels are the name of the sustainability certification, "other", "unknown", and "unavailable". The target variable value, certification status, is true when the product has at least one certification named.

4.3 Data Characteristics

GreenDB contains more than 2,5 million clothing items classified by a single category. The majority of products, 80 %, comes from the German market. 14 % comes from the French market and the remaining 6 % comes from the British market.

The majority of brands in the whole GreenDB dataset had 10 % or less of their products certified. This shows that the product's certification status is not directly dependent from the brand.

Of all the clothing items, 76 % does not have a certification. 23 % has one certification (other than "unknown", "unavailable", or "other"). And the rest of the clothing items have from 2 to 5 certifications. These different frequencies are shown in the figure 4.1.

4.4 Logistic Regression

Logistic regression was chosen as the model as it is a simple and interpretable model that balances the use of real-life unlabeled data, which is not ideal for training a model to detect a signal. Real-life data is often unbalanced with no certainty

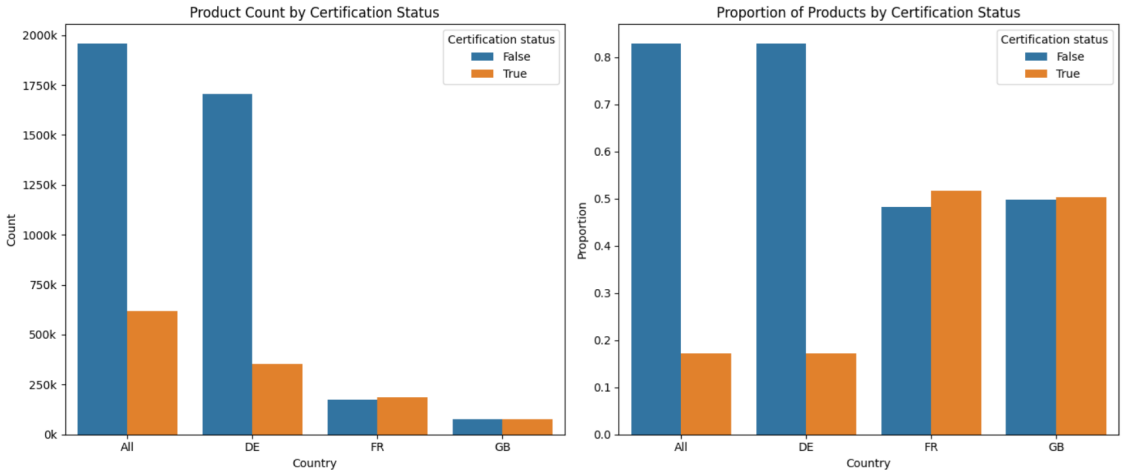


Figure 4.1: Product Count and Proportions of Certified Products

of containing a signal that can be predicted, which is the case of purpose fitted training data. Therefore choosing a simple model is appropriate for predicting the certification status as it is less likely to overfit and is easier to interpret, which is important with unbalanced data. The task of predicting whether or not a product has a certification or not is a binary classification task, where the prediction is a probability value between zero and one.

Logistic regression was chosen for the prediction of the certification status as it is a simple model that is well adapted to the task of predicting binary probabilities. It is a statistical method that focuses on the relative probability (odds ratio) of an event occurring versus not occurring, taking into account the weight of each feature [30]. It explains how the learning data's binary predictors affect the dichotomous dependent target variable by calculating a probability value. The weights of the logistic regression model can be interpreted as the magnitude of the effects of the variables on the outcome.

Logistic regression predicts the probability of the positive class $P(\hat{y}_i = 1 \mid \boldsymbol{x}_i)$ as

$$\hat{p}(\boldsymbol{x}_i) = \frac{1}{1 + \exp(-\boldsymbol{x}_i \boldsymbol{w})}$$

where $\boldsymbol{x}_i$ is the input feature vector (the vector of generic environmental claims detected) for the ith observation, and $\boldsymbol{w}$ denotes the weight for each feature. If $P(\hat{y}_i = 1 \mid \mathbf{x}_i) \geq 0.5$, the product's certification status is classified as true. [30]

The prediction of certification status can also be used to investigate and gain information on the possible existence of greenwashing. The performance of the logistic regression can be seen as an indicator of the effects of suspected greenwashing, which might weaken the prediction results. If the generic environmental claims present in the product descriptions exist only within certified products, as is the goal of the directive (EU) 2024/825, the mere existence of such a term in the description should systematically translate to a high probability of a positive target value, meaning that the product should always be certified if the environmental claim is justifiable.

Logistic regression is a statistical method focusing on the relative probability (odds ratio) of an event occurring versus not occurring. [30] This probabilistic binary nature makes it suitable for the classification task of predicting a product's certification status based on the detected terms in the product description.

4.5 Course of the Study

The application part of the thesis aimed to find out if product descriptions can be used in predicting the product's sustainability, depicted by the product having at least one certification. The challenges of real-life data proved to be a significant obstacle, and thus in order to obtain better results, several different models were used for predicting the certification status. Every time that the model showed to be insufficient for producing any reliable results, a new one was constructed with a simpler logic.

In each test setting product descriptions were combined with a logistic regression model trained on training data to predict the product's certification status in test data. The different models started from automatic selection of the most important

words for prediction with TF-IDF. Then generic green claims present in the product descriptions were used as predictors of the sustainability status. Models using generic green claims were made gradually simpler and combined with sustainability claims as features. Finally the relationship between a generic sustainability claim and the certification status was determined with odds ratios, by combining all the given claims into a single indicator variable.

The course of the study:

1. TF-IDF
2. Generic environmental claims with logistic regression
3. Generic sustainability claims as term groups with logistic regression
4. Stratified analysis with odds ratios

4.5.1 TF-IDF

The first test setting was done with Term Frequency - Inverse Document Frequency transform (TF-IDF) together with logistic regression as an automated way to identify the most important terms for predicting the product's certification status. The scikitlearn's TfidfVectorizer was used, that converts a collection of raw documents to a matrix of TF-IDF features [46].

Qaiser and Ali (2018) depict TF-IDF as a numerical statistic that shows the relevance of keywords to a specific document. It consists of two parts: the term frequency measures how many times a term is present in the document, and the inverse document frequency assigns lower weight to frequent words and greater importance to infrequent words. [47] TF-IDF, the multiplication of term frequency and inverse document frequency, makes a weighted calculation of the most important terms in the documents for determining the outcome.

Unfortunately it turned out that the amount of products in each subset was too large to find relevant differences between terms, as enough supporting evidence was found for any arbitrary term. This was shown in the rather good evaluation results for completely arbitrary terms chosen from the 1000 most important terms given to the logistic regression model.

For example the ten most importand terms with their coefficients in the British subset were:

- marc: 7.657
- ups: 6.657
- boots: 6.062
- 2023: 6.051
- ons: 5.862
- belt: 5.198
- ballet: 5.063
- sandals: 5.024
- leather: 4.931
- gabor: 4.905

It seems that some of the terms found refer to brands or material, which might explain the relation with sustainability, but other terms seem to be completely random.

4.5.2 Features: Generic Environmental Claims

Investigating the relationship between sustainability and green marketing requires the selection and identification of environmental claims, and researching their relationship to a variable that depicts sustainability. Environmental claims can be any claims that refer to any of the assumed sustainability and environmental impact characteristics of a product, which results in a vast amount of possible green claims.

The directive (EU) 2024/825 prohibits "the making of a generic environmental claim without recognised excellent environmental performance which is relevant to the claim" as well as "generic claim such as ‘conscious’, ‘sustainable’ or ‘responsible’ based exclusively on recognised excellent environmental performance, because such claims relate to other characteristics in addition to environmental characteristics, such as social characteristics" [2]. The examples of generic environmental claims given in the directive were used as the second model’s features for predicting the certification status. Using the same terms in each three language areas would have provided a way to compare the terms used in each market and their importance for the prediction, had there been enough matches between the generic environmental claims and the product descriptions.

4.5.3 Data Preparation

As the GreenDB data comes from real webshops, the data used in creating the logistic regression models and finding generic environmental claims in product descriptions had to be prepared before being used in a logistic regression model. The preprocessing was necessary as the generic environmental claims consisted of several words and were subject of inflection: changes expressing different grammatical categories like plurality or gender. Both the terms and decriptions had to be preprocessed in the same way in order to find matches between the two sets.

Using the generic environmental claims as features for the logistic regression

model is a form of natural language processing. Natural language processing requires special task-specific data preparation. The most important forms of preparation for textual data is removing stopwords, e.g. "a", "and", "if", and normalizing the text by removing special characters like quotation marks or punctuation.

Finding matches between terms consisting of one or several words and a mass of text in the product description is a task where the inflection of words has to be taken into account. Manipulating each word by dropping its prefixes or suffixes leaves only the root of the word called the stem. Stemming the words enables matching the words in a text with their tokens despite the inflection. Stemming was done both to product descriptions and the prediction feature terms with Natural Language Toolkit's Snowballstemmer before the feature terms could be detected in the product descriptions.

After the stemming, a dictionary was created of the generic environmental claims to map the original terms with their stemmed versions. Matches between the tokens in the dictionary and the product description were then saved into a list containing all the matched terms between the generic environmental claims and the product's description. With Scikit-learn's MultiLabelBinarizer the list of terms was transformed into a one-hot-encoded dataframe of the generic environmental claims found in the product descriptions. A generic environmental claim had its own column, with the value one if it had any matches in the product description, and zero if no matches were found.

4.5.4 One-Hot-Encoded Generic Environmental Claims

In the second test setting the generic environmental claims from the directive (EU) 2024/825 were used. From the matches between generic green claims and product decsriptions, ten most common were chosen and turned into one-hot-encoded vectors, acting as the features of the logistic regression model.

The list of generic environmental claims proposed in the directive (EU) 2024/825 [2]:

- Great Britain: ‘environmentally friendly’, ‘eco-friendly’, ‘green’, ‘nature’s friend’, ‘ecological’, ‘environmentally correct’, ‘climate friendly’, ‘gentle on the environment’, ‘carbon friendly’, ‘ energy efficient’, ‘biodegradable’, ‘biobased’, ‘conscious’, ‘sustainable’ ‘responsible’, ‘made with recycled material’, ‘climate neutral’, ‘CO2 neutral certified’, ‘carbon positive’, ‘climate net zero’, ‘climate compensated’, ‘reduced climate impact’, ‘limited CO2 footprint’.

- France: «respectueux de l’environnement», «respectueux de la nature», «vert», «ami de la nature», «écologique», «bon pour l’environnement», «bon pour le climat», «favorable à l’environnement», «à faible intensité de carbone», «économe en énergie», «biodégradable», «biosourcé», «respectueux», «durable», «responsable», «fabriqué avec des matériaux recyclés», «neutre pour le climat», «certifié neutre en CO2», «bilan carbone positif», «zéro net pour le climat», «climatiquement compensé», «impact réduit sur le climat», «empreinte CO2 limitée».

- Germany: „umweltfreundlich“, „umweltschonend“, „grün“, „naturfreundlich“, „ökologisch“, „umweltgerecht“, „klimafreundlich“, „umweltverträglich“, „CO2-freundlich“, „energieeffizient“„biologisch abbaubar“, „biobasiert“, „bewusst“, „nachhaltig“ oder „verantwortungsbewusst“, „mit Recyclingmaterial hergestellt“, „klimaneutral“, „zertifiziert CO2-neutral“, „CO2-positiv“, „mit Klimaausgleich“, „klimaschonend“, „mit reduziertem CO2-Fußabdruck“.

For each subset, ten of the most common terms were chosen for the logistic regression model:

- Great Britain: "sustainable", "green", "responsible", "conscious", "eco-friendly",

"ecological", "environmentally friendly", "biodegradable", "made with recycled material", "climate neutral"

- France: «durable», «responsable», «vert», «écologique», «respectueux», «respectueux de l'environnement», «biodégradable», «fabriqué avec des matériaux recyclés», «bon pour l'environnement», «bon pour le climat».

- Germany: „nachhaltig“, „ökologisch“, „grün“, „umweltfreundlich“, „umweltschonend“, „bewusst“, „klimaneutral“, „verantwortungsbewusst“, „biologisch abbaubar“, „umweltverträglich“.

Their frequencies are shown in figure 4.2.

With the terms taken from the directive (EU) 2024/825, only a small part of the product descriptions were matched per subset.

The top ten most frequent generic environmental claims were largely the same in the three subsets. The most frequent terms were "sustainable", together with "green", "ecological", and "environmentally friendly".

Regarding the sets of the 10 most frequent claims, almost all of the same terms were present in each subset, indicating their importance and the similitude between markets. Concerning the other generic environmental claims, the cultural differences in use of vocabulary were more visible.

The terms in the British subset were distributed quite evenly, but in the French and German subsets one or two terms prevailed significantly, while the others were present in fewer cases. For example in the French subset, the term "durable" occurred in 25 000 descriptions more than the next used term "responsable".

All in all, the amount of matches was not large enough to produce reliable results. In each subset, less than 10 % of the product descriptions contained any of the given terms, which was not enough to produce any relevant information on the relation between the terms and certifications. This could be seen in the linear ROC-curves

and histogram of predicted values, all concentrated around 0.5.

Generic Sustainability Claims as Word Groups

As the ten most common generic environmental claims were insufficient for providing a strong enough signal for predicting the certification status, the next model was done by using the same generic environmental claims, completed with other similar generic sustainability claims. These claims were then combined into four groups: words regarding environment, climate, sustainability, and material. The four groups of generic sustainability claims were then one-hot-encoded into four predicting features for the model. For example if the description contained any of the words "organic", "bio", "recycled", "biodegradable", or "biobased", value one was given for the material terms column.

Combining the generic sustainability terms into four groups improved the prediction power of the model slightly, as the number of matches between grouped terms grew slightly. The amount of matches was still not sufficient for creating reliable results, keeping the model's prediction power at the level of random guessing.

Based on the coefficients, the most important term group for predicting the certification status in each group was climate related generic sustainability claims, even though it had the least amount of matches in product descriptions of all the word groups.

4.5.5 Stratified Analysis with Odds Ratios

As the amount of generic sustainability claims was not sufficient for logistic regression, statistical testing was used to determine if there is a statistically significant relationship between a generic sustainability claim and the certification status. The generic sustainability claims used in the last logistic regression model were combined into a new binary variable, that indicates if a product description contains a generic

sustainability claim or not. Based on this new variable depicting exposure, denoted by Y, and the outcome variable certification status, denoted by X, 2×2 contingency tables of the sample frequencies were created. In the these tables, the binary variable categories 1 and 2 correspond the actual values 1 and 0 in the data. Category 1 represents an actual value 1 and category 2 represents an actual value 0.

Contingency Tables

According to Andrés et al. (2015), frequencies set into 2×2 tables are used to test if association exists between two dichotomic qualities. The rows contain two levels of one of the qualities, the columns contain the two levels of the other quality, and inside the table are the observed frequencies. [48] Hence each cell of the contingency table has characteristics of both Y and X, denoted by n_{ij}.

According to Kateri (2014), the hypothesis of interest in a cross-classified table of two binary variables on a sample is the independence of X and Y classification variables. In a contingency table, n_{ij} denotes the observed cell frequency at cell (i, j), and the "+" in place of the index denotes the summation over the given index. [49]

	$X = 1$	$X = 2$	n_{i+}
$Y = 1$	n_{11}	n_{12}	n_{1+}
$Y = 2$	n_{21}	n_{22}	n_{2+}
n_{+j}	n_{+1}	n_{+2}	n

With the aforementioned variables X and Y, $X = 1$ refers to a certified product with variable value 1 and $X = 2$ to an uncertified product with variable value 0. $Y = 1$ refers to an existing generic sustainability claim in the product description with variable value 1 and $Y = 2$ to no matched claims with variable value 0.

Stratification

Differences between markets were visible both in the frequency of the generic sustainability claims and the terms used, which might lead to confounding. According to Rothman et al. (2024), confounding comes from the mixing of the effect of the confounding variable with the effect of the exposure. Confounding occurs because the comparison of the two groups is also a comparison of differing distributions of the confounding factor. [50] According to Webb et al. (2024), a confounder is a factor that is associated with both the exposure and the outcome, that is either causal or a proxy of the true cause. The confounder must not be a consequence of either the exposure or the outcome. [51]

According to Tripepi et al. (2017), to obtain an unbiased estimate of the causal relationship between exposure and outcome, confounding must be controlled. The simplest method to control confounding during data analysis is stratification. It allows to control confounding by creating subgroups in which the confounding variable does not vary [52]. According to Andrés et al. (2015), the aim of stratification based on a covariate is to contrast the independence of both the original dichotomic qualities, taking into account the heterogeneity of the populations defined by strata [48]. Stratification also provides a simple and direct way to evaluate the effect-measure modification, meaning how the relationship between e.g. an exposure and disease depends on some third variable [50]. The products were hence stratified into three groups based on the market area covariate, producing three stratified 2×2 contingency tables 4.1, 4.2, and 4.3 along with the total data 2×2 table 4.4. This test setting shows if there are differences between the market areas [48].

Table 4.1: Stratified British 2×2 Contingency Table with Totals

Claim status	Certified	Uncertified	Total
Claim found	14 490	7 358	21 848
No claims	62 165	68 599	130 764
Total	76 655	75 957	152 612

Table 4.2: Stratified French 2 × 2 Contingency Table with Totals

Claim status	Certified	Uncertified	Total
Claim found	28 457	20 289	48 746
No claims	158 525	154 114	312 639
Total	186 982	174 403	361 385

Table 4.3: Stratified German 2 × 2 Contingency Table with Totals

Claim status	Certified	Uncertified	Total
Claim found	62 289	193 968	256 257
No claims	291 274	1 511 977	1 803 251
Total	353 563	1 705 945	2 059 508

Tripepi et al. (2017) depict the steps of assessing confounding through the Mantel-Haenszel formula:

1. Calculate the crude odds ratio θ (i.e. without stratifying)
2. Stratify by the confounding variable and calculate stratum-specific θ
3. Assess the homogeneity of the effect estimates across strata and compare stratified and unstratified θ
4. If there is homogeneity in effect est imates across strata then calculate the overall, adjusted θ by the Mantel-Haenszel formula
5. If there is heterogeneity and we are interested in effect modification, stratum-specific effect estimates should be reported separately

[52]

Calculating Odds Ratio from a Contingency Table

According to Kateri (2014) the results of a binary response are often presented and interpreted not directly on the success probability π but regarding success's relative importance to failure. The key quantity is the odds of success, meaning the ratio of

Table 4.4: Whole data 2×2 Contingency Table with Totals

Claim status	Certified	Uncertified	Total
Claim found	105 236	221 615	326 851
No claims	511 964	1 734 690	2 246 654
Total	617 200	1 956 305	2 573 505

success vs. failure probabilities for an outcome

$$odds = \frac{\pi}{1-\pi}$$

[49]. For calculating the odds, conditional probabilities are needed. For calculating the probabilities, contingency tables are used.

A contingency table from randomly sampled units will have a multinomial distribution with a parameter vector $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = \pi_{ij}$ where $\pi_{ij} = P(Y = i, X = j)$ is the joint probability that a randomly selected individual falls into the (i, j)th cell of the contingency table [53]. This means that the probability vector $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ is the joint distribution of the classification variables X and Y [49].

Summing the joint probabilities over one variable gives the marginal distribution π_{i+}, for example $P(Y = i) = \pi_{1+}$. The observed marginal distribution of Y is $p_{i+} = \frac{n_{i+}}{n}$, for example $p_{1+} = \frac{n_{1+}}{n}$. [53]

The probability of the ith row category is $P(Y = i) = \pi_{i1} + \pi_{i2} = \pi_{i+}, \quad i \in \{1, 2\}$ and the probability of the jth column category is $P(X = j) = \pi_{1j} + \pi_{2j} = \pi_{+j}, \quad j \in \{1, 2\}$. In matrix notation this makes

π_{11}	π_{12}	π_{1+}
π_{21}	π_{22}	π_{2+}
π_{+1}	π_{+2}	$\pi_{++} = 1$

The underlying probability pattern of the 2×2 contingency table formed by two

independent binomials regarding the exposure variable is

$$\begin{array}{cc|c} \pi_1 & 1-\pi_1 & 1 \\ \pi_2 & 1-\pi_2 & 1 \end{array}$$

and the basic associated hypothesis testing problem is $H_0 : \pi_1 = \pi_2 (= \pi)$. [49]

For observed data, p can be used instead of $\hat{\pi}$ to represent a sample proportion, where $p_{i,j} = \frac{n_{ij}}{n}$ is the sample proportion of observation in the (i,j)th cell [53]. Hence the probabilities can be directly calculated from the sample proportions in a 2×2 contingency table, where the frequencies represent the sample probabilities of the product having a given quality; being certified or not or having a generic sustainability claim in the product description.

The conditional probability distribution is a probability of one variable given the values of the other variable. For the variable X, given values of Y, the conditional distribution is $\pi_{j|Y=i} = \frac{\pi_{ij}}{\pi_{i+}}$, such that $\sum_j \pi_{j|Y=i} = 1$. The conditional distribution depicts how the distribution of X changes as the categories of Y change. [53]

The conditional probability of a certified product given an existing generic sustainability claim is calculated as

$$P(X = 1|Y = 1) = \frac{\pi_{11}}{\pi_{11} + \pi_{12}} = \frac{\pi_{11}}{\pi_{1+}}$$

The observed conditional probability distributions of X, given Y can be calculated directly from the contingency table by

	$X = 1$	$X = 2$	total
$Y = 1$	$\frac{n_{11}}{n_{1+}} = p_{1\|Y=1}$	$\frac{n_{12}}{n_{1+}} = p_{2\|Y=1}$	1
$Y = 2$	$\frac{n_{21}}{n_{2+}} = p_{1\|Y=2}$	$\frac{n_{22}}{n_{2+}} = p_{2\|Y=2}$	1

[53]

The odds of an occurrence $\frac{p}{1-p}$ is the ratio of the probability of an outcome to the probability of the opposite outcome within the determined group of exposure, calculated from the conditional probabilities by dividing the probability of an occurrence by the probability that it does not happen. To get the odds of a certified product given an existing generic sustainability claim, the conditional probability of a certified product given an existing generic sustainability claim is divided by the conditional probability of an uncertified product given an existing generic sustainability claim:

$$Odds(X = 1|Y = 1) = \frac{P(X = 1|Y = 1)}{1 - P(X = 1|Y = 1)} = \frac{P(X = 1|Y = 1)}{P(X = 2|Y = 1)}$$

From the contingency table the estimated odds of a certified product given an existing generic sustainability claim is then:

$$o_{1|Y=1} = \frac{n_{11}/n_{1+}}{n_{12}/n_{1+}} = \frac{n_{11}}{n_{12}}$$

The odds ratio θ is the ratio of the odds under one set of conditions to the odds under another set of conditions [54]. It is more informative for the comparison of the success probabilities π_1 and π_2 than their difference, as it incorporates the relative importance of success probabilities in terms of their level of magnitude [49].

Due to the conditional probability

$$\pi_i = P(X = 1|Y = i) = \pi_{1|i}, \quad i \in \{1, 2\}$$

where the probability of success is dependent on the explanatory variable, the odds ratio can be equivalently defined in terms of the joint distribution of a 2×2 contingency table as

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

and the sample odds ratio as

$$\hat{\theta}(\mathbf{n}) = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Therefore in the 2×2 contingency tables the variable Y, which indicates the existence of a generic sustainability claim in the product description, is cross-classified with the outcome X (success-failure). [49]

Hence the odds ratio θ of a certified product relative to having a generic sustainability claim in the product description is the odds of a certified product given an existing generic environmental claim in the product description, divided by the odds of a certified product given a non-existing generic sustainability claim in the product description.

Stratum-specific odds ratios are calculated within each stratum of the confounding variable and compared with the corresponding odds ratios in the whole group. When there is no confounding, the odds ratios are roughly homogenous across strata and do not differ from that in the whole group. [52] Each stratum provides a separate, stratum-specific estimate, of the same overall effect. [50] The stratum-specific odds ratios are listed in table 4.5.

Table 4.5: Key Figures of a Product's Certification Status by Generic Sustainability Claim Status by Market Area

Key figure	British	French	German	Total
$P(X = 1\|Y = 1)$	0,66	0,58	0,24	0,32
$Odds(X = 1\|Y = 1)$	1,97	1,4	0,32	0,47
θ	2,17	1,36	1,67	1,61
θ 95% CI	[2,11; 2,24]	[1,34; 1,39]	[1,65; 1,68]	[1,60; 1,62]

Note: Category X=1 indicates a certified product and category Y=1 indicates at least one generic sustainability claim found in the product description.

Mantel-Haenszel Test

According to Ufondu et al. (2023), odds ratios from case-control studies can be estimated with the Mantel-Haenszel test, by doing a stratified analysis involving a 2×2 contingency table. It is an ideal test to use when controlling for a confounding variable. The Mantel-Haenszel test can be used to confirm that there is no association between the confounding variable and independent variable outside of the dependent variable. A single, weighted summary of association, the average of the odds ratio, is produced by the test. [55]

According to Vierra et al. (2023), the Mantel-Haenszel test is uniformly the most powerful and unbiased analytical test to utilize when there is a constant odds ratio among each of the two by two tables under study [56]. Andrés et al. (2015) claim that "The Mantel-Haenszel test is the most frequent asymptotic test used for analyzing stratified 2×2 tables." [48].

In a stratified analysis like the Mantel-Haenszel, the result is a single summary of the relation between exposure and outcome over strata. It compares exposed and unexposed subjects within each stratum and then aggregates the information from these stratum-specific comparisons across all the strata by taking a weighted average. [50]

The Mantel-Haenszel formula allows to calculate an overall, unconfounded, adjusted, effect estimate of a given exposure for a specific outcome by pooling stratum-specific odds ratios [52]. The adjusted odds ratio is a weighted average of the stratum-specific odds ratios, where the influence of sample size, and thus precision, is taken into account. It summarizes the effect of the exposure for the confounder. [51]

The weight for each stratum is

$$w = \frac{n_{12} \times n_{21}}{n}$$

Each individual value is multiplied by its weight, added up and divided by the sum of the weights, creating a pooled odds ratio. [51] Pooling is a method for obtaining unconfounded estimates of an effect measure across a set of strata, where a weighted average of the stratum-specific estimates of effect is taken. The weights are assigned so that the strata that provide the most information (the most data), get the most weight. [50] The commonly used weights for pooling odds ratios are the ones proposed by Mantel and Haenszel. [51]

For each stratum the odds ratio is multiplied by weight,

$$\theta \times w = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}} \times \frac{n_{12} \times n_{21}}{n} = \frac{n_{11} \times n_{22}}{n}$$

summed over all strata, and divided by the sum of the weights in order to get the Mantel-Haenszel pooled odds ratio:

$$\text{Mantel–Haenszel pooled } \theta = \frac{\sum [(n_{11} \times n_{22}) \div n]}{\sum [(n_{12} \times n_{21}) \div n]}$$

[51]. The Mantel-Haenszel pooled odds ratio for certification status by generic sustainability claim status for all market areas together with the confidence interval of 95%, p-value and Breslow-Day p-value is listed in table 4.6.

Table 4.6: Mantel–Haenszel Pooled Odds Ratio

MH pooled θ	CI 95%	p-value	Breslow-Day p-value
1,64	[1,62; 1,65]	0.00	0.00

4.6 Results

An odds ratio greater than one refers to the generic sustainability claim having a positive impact on the certification status. For the entire data set, the odds of being certified are 1,6 times higher for products with at least one generic sustainability

claim in the product's description, than for a product without generic sustainability claims as seen in table 4.5.

The differences in subsets show the differences between the market areas. The British products had a clearer difference in certification status between the products that had a generic sustainability claim and those that didn't, and the French products have the smallest difference.

British products with a generic sustainability claim in their product description have more than twice the odds of being certified compared to the British products without a generic sustainability claim. French products with a generic sustainability claim have 36 % higher odds of being certified, than the French products without a generic sustainability claim. For German products, the odds are 66 % higher.

The association between a generic sustainability claim and a certification was weaker in the French subset, which might be due to a weaker affiliation or explainable by the large quantity of mathes for the term 'durable', which might not always be used to depict sustainability. The same odds ratio for the entire data set and the German data is probably due to the German data making up a large proportion of the entire data set.

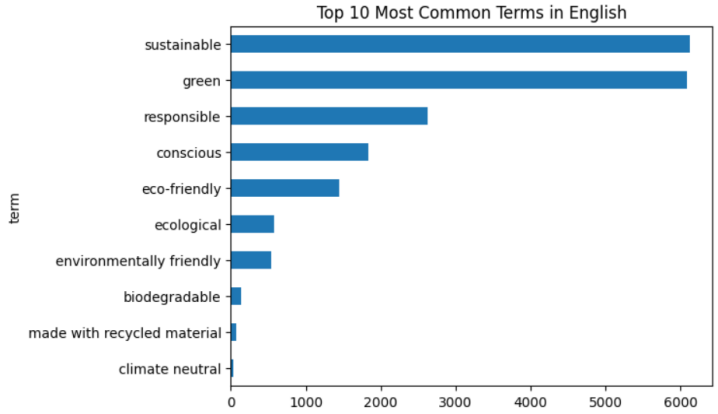
Based on the odds ratios, the presence of a generic sustainability claim can be used as a predictor of a certified product, and that the predictive power depends on the market region. The strongest affiliation is in Great Britain and the weakest in France. The narrow confidence intervals show that the odds ratios are precise with very slight uncertainty.

The stratified Mantel-Haenszel odds ratio of 1.64 means that having at least one generic sustainability claim in the product description increases the odds of the product having a certification by 64 %, also when taking into account the differences in market regions.

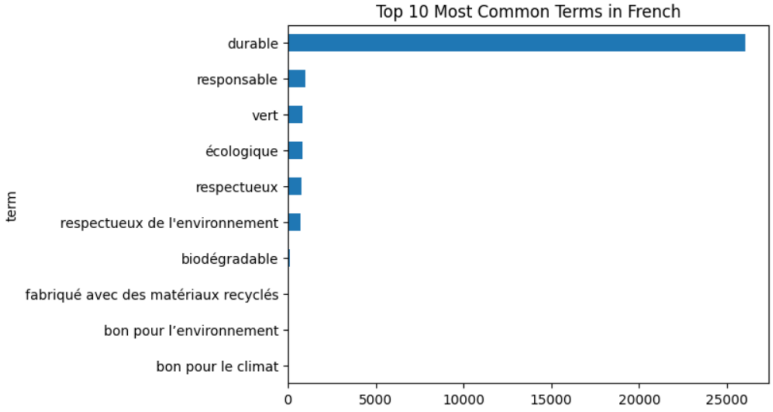
The p-value of the Mantel-Hazensel test shows that the hypothesis of the pooled

odds ratio being equal to 0 is statistically extremely unlikely. The Breslow-Day test's p-value, for the hypothesis that all odds ratios are identical, shows that the odds ratios across strata are not similar and that the stratification is useful.

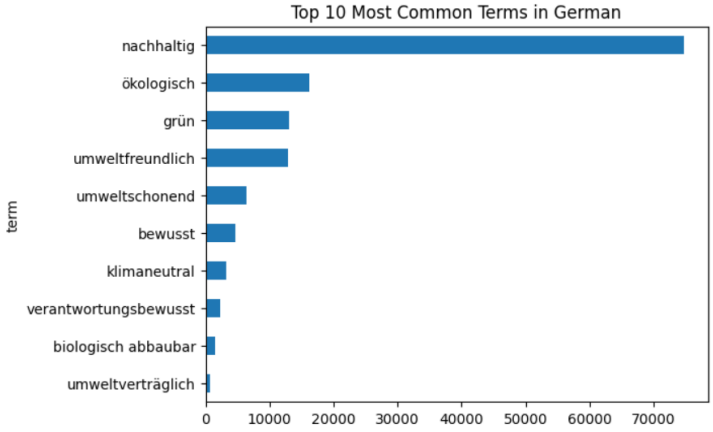
Figure 4.2: Frequencies of the Generic Environmental Claims in the Three Market Areas



(a) Frequencies of the English terms



(b) Frequencies of the French terms



(c) Frequencies of the German terms

5 Conclusions and Future Research

The inspiration for the thesis came from the directive (EU) 2024/825 which aims to enable customers to assess the sustainability of a product more reliably. While shopping in a webshop, the customer is faced with very limited information about sustainablility and the credibility of the marketing claims. With the rapid development of various machine learning methods combined with a growing urgency to shop sustainably, the risk and characteristics of greenwashing can be assessed with data analytics and the relationship between the sustainability related terms and their credibility can be described in more ways than before.

The goal of this thesis was to research sustainability and green marketing by describing the relationship between the sustainability claims of the product description and the certification status of the product. As a part of this thesis a logistic regression models that could predict if a product has at least one certification based on the sustainability claims used in the product description were made but with no valid predictions. The models were limited by the sparse features of the real-life data, the ambiguity of sustainability and the problem of reliably verifying the sustainability of the product and hence the justification of the generic sustainability claims. Because of these challenges, it was not possible to state if greenwashing prevented the prediction of the certification status, and on the other hand, greenwashing is likely to have prevented the robust prediction of the certification status.

The results of the odds ratios and stratified Mantel-Haenszel test showed that

there is a statistically significant relationship between a generic sustainability claim present in the product description and the certification status of the product.

The research question was "How strongly is a sustainability claim of a product description associated with the product's certification status?". The answer is that in real-life webshop data the association is statistically significant but not strong enough to be a reliable indication of real sustainability of the product. The restrictions of real-life data, challenges of measuring sustainability, and possible effects of greenwashing made it hard to answer the research question. As a conclusion based on the results on webshop data on clothing, having at least one generic sustainability claim in the product description increases the odds of the product having a certification by 64 % but a consumer cannot rely strictly on the sustainability claims made in product descriptions and needs more information to support the sustainability claims.

5.1 Discussion

The results of multiple different test settings show that reliable conclusions are very hard to draw from real-life data. The major issue is the extreme sparsity in the model's features, which leads to all predictions produced by the logistic regression models being around 0,5, resulting in negative bias of the model as the treshold value is not fulfilled.

The other problem in researching sustainability claims is the effects of greenwashing, where unsustainable products are passed as sustainable. This prevents the use of sustainability claims as direct indicators of the product's sustainability, as part of the sustainability claims are not associated with a ceritified product. Greenwashing combined with the nature of real-life data, where the descriptions are mostly short, and the sustainability claims used as a predictor were sparse, meant that a clear signal necessary for building a strong model was missing.

The difficulties produced by the real-life data on the model building resulted in building a sequence of models from more elaborated to simpler, as the models had to be simplified each time when the previous version did not provide sufficient results due to the aforementioned problems in order to try and gain more matches between the generic sustainability claims used as model features and the product descriptions.

The logistic regression models used generic sustainability claims found in product decriptions as labels for predicting the certification status of a clothing product. If only certified products used these terms in their descriptions, predicting the certification status should be an easy task. However combined with possible greenwashing, where the same environmental or sustainability terms are used by certified products and uncertified products without justification, the classification task becomes more complicated.

The logistic regression models were not able to produce any reliable results. This is probably due to the extremely sparse feature matrix where only a small portion of the products contained the environmental claims, which in turn did not provide a strong enough signal for reliable prediction. In each subset, around 10 % or less of the product descriptions contained any of the general green claims, resulting in 90 % of the data not containing the signal and making the classification task more complicated. The problem of sparse features means that the models did not improve with optimization, regularization tuning or TF-IDF, that were tested during development. With each modification the models offered arbitrary results at best. At the same time some of the uncertified products did contain the generic environmental claims tested, indicating a possibility of greenwashing.

5.2 Limitations

Using descriptive data analytics to provide insight to the relationship between sustainability claims and the certification status was complicated because of the the extremely sparse training data. GreenDB dataset has a very limited number of products whose description contained any of the terms given in the directive (EU) 2024/825 or the other generic sustainability claims added to the other models. This affected the binary prediction of the models by setting the probability of the certification status as 0,5 each time. The problem of biased data stems from the fact that GreenDB's data is real-life data taken directly from real webshops. Training the model with a data set that contains labeled examples of greenwashing would be significantly easier.

However according to Calamai et al. (2025), there is currently no large-scale dataset of instances of greenwashing. Due to the lack of dataset of instances of greenwashing with positive and negative examples and hence most existing work targeting intermediate tasks such as identifying climate-relatedness or green claims instead of identifying greenwashing as a whole, the detection of greenwashing has remained on a theoretical level without empirical validation. [44] There are still ways to work around the problem of data, as Kobti et al. (2021) did. As greenwashing is a common marketing practice in many domains, training a classifier with real examples on all possible domains is challenging, hence an easier approach would be to pretrain the model on one domain and generalize it on another [35].

Closer considerations also revealed that there seems to be inconsistencies between listed certifications and certifications claimed in the product description, which weakens the quality of the predictive model. A great number of product certifications were listed as "unknown" and "unavailable", which is not reliable information.

The ideal situation would be a link between GreenDB and a database containing

third party evaluations of the sustainability of the same brands or products. This would allow for a more robust prediction. For this purpose, Good on You, Transparency Index, Know the Chain, and Textile Exhange were considered but none had a sufficient overlap of brands with GreenDB to reliably train a logistic regression model.

Detecting greenwashing contains several other problems. According to Calamai et al. (2025), the three fundamental challenges are the absence of a commonly accepted and actionable definition of greenwashing that creates ambiguity, significant legal and reputational consequences of being accused of greenwashing, and the lack of explicit precedents that could serve as jurisprudence for systematic labeling, which makes the annotation of actual examples of greenwashing challenging. [44]

5.3 Future Research Directions

There are several indications for development and future research that can be drawn from the data analytics approach used, combined with previous research on the same subject.

Based on results of the descriptive data analytics and odds ratios, further investigation on the relationship between sustainability claims and certification status should be done using different methods to gain further insight into the role of greenwashing as a possible explanation of the poor performance on the predictability of the certification status and the predicting power of the sustainability claims, as well as experiments on another set of data, preferably a labeled data set.

The biggest obstacle seems to be the real-life data, with extreme sparsity in the features that are used to predict the certification status. Ideally in order to research the effects of greenwashing on the relation between sustainability claims and sustainability status, a labelled training set with verified instances is required to get reliable outcomes. If not labelled, the data should be at least balanced with

sufficient features to provide a reliable prediction.

Another important obstacle to solve is the target variable, that proves a product's sustainability. If there is not a strong enough signal between the sustainability claims and certificate status, another variable should be used instead. One option for keeping the certification status as the target variable would be ranking the type of the certification of certified products. The subject of research could then be the relationship between green marketing and the strictness of the certification. This would require research on the criteria of the certification and its compliance with the ISO 17065 standard, as stated in the directive (EU) 2024/825.

Ideally there would be a different database of third party assessments of the product or brand level sustainability, that could be combined with GreenDB with enough coverage to have enough data for training and test sets.

Calamai et al. (2025) give detailed insights for further research on developing machine learning models for greenwashing identification. According to them, problems in previous research on identifying greenwashing with machine learning are evaluation methodology, model robustness, quantifying information and relying on regulatory claims. In further development, correct evaluation of complete tasks can include building robust evaluation methods for generated texts or building classification benchmarks that are difficult to solve but simple to evaluate. In practice this could mean hyper-parameter tuning, simple baselines, and inter-annotator agreement measures. Model evaluation should also include quantifying the uncertainty of the performance measures by confidence intervalls or robustness assessments, with real-world applications that exhibit data quality issues or highly imbalanced classes. Instead of complying only the text related to a topic, also the information itself should be quantified. In addition to identifying supposed characteristics of a misleading claim, the statements should be confronted to regulatory texts as they are becoming more precise about the requirements of environmental communications.

[44]

As Calamai et al. (2025) state, previous research on understanding the characteristics and formalizing greenwashing has created the basis for developing AI systems capable of automatically identifying these misleading communications. From theoretical understandning of greenwashing and the detection of signals that can indicate misleading climate-related communications, future research should move towards building a representative and diverse corpus of real-world cases of potential greenwashing to first evaluate models and, secondly, to empirically analyze its mechanisms and patterns. [44]

References

[1] A. Persakis, T. Nikolopoulos, I. C. Negkakis, and A. Pavlopoulos, "Greenwashing in marketing: A systematic literature review and bibliometric analysis", *International Review on Public and Nonprofit Marketing*, vol. 22, no. 4, pp. 957–992, 2025.

[2] European Parliament, *Directive (EU) 2024/825 of the European Parliament and of the council, Of 28 february 2024 amending directives 2005/29/EC and 2011/83/EU as regards empowering consumers for the green transition through better protection against unfair practices and through better information*, 2024. [Online]. Available: `https://eur-lex.europa.eu/eli/dir/2024/825/oj/eng`.

[3] W. Moodaley and A. Telukdarie, "Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review", *Sustainability*, vol. 15, no. 2, 2023.

[4] U. Nations, *Sustainability*. [Online]. Available: `https://www.un.org/en/academic-impact/sustainability`.

[5] G. Stöckigt, J. Schiebener, and M. Brand, "Providing sustainability information in shopping situations contributes to sustainable decision making: An empirical study with choice-based conjoint analyses", *Journal of Retailing and Consumer Services*, vol. 43, pp. 188–199, 2018.

[6] E. Kesidou and C. Palm, "Eco-credentials in the fashion and textile industry: Assessment and evaluation: A review of eco-credentials, their strengths and weaknesses, and recommendations for improvement", *A Research Report by University of Leeds Business School for the Back to Baselines in Circular Fashion & Textiles*, 2024. [Online]. Available: `https://ssrn.com/abstract=5052003`.

[7] D. O'Rourke and R. Abraham, "The impact of sustainability information on consumer decision making", *Journal of Industrial Ecology*, vol. 20, no. 4, pp. 882–892, 2016.

[8] M. Wojnarowska, M. Sołtysik, and A. Prusak, "Impact of eco-labelling on the implementation of sustainable production and consumption", *Environmental Impact Assessment Review*, vol. 86, 2021.

[9] S. Plakantonaki et al., "A review of sustainability standards and ecolabeling in the textile industry", *Sustainability*, vol. 15, no. 15, 2023.

[10] J. de Boer, "Sustainability labelling schemes: The logic of their claims and their functions for stakeholders", *Business Strategy and the Environment*, vol. 12, no. 4, pp. 254–264, 2003.

[11] M. Yokessa and S. Marette, "A review of eco-labels and their economic impact", *International Review of Environmental and Resource Economics*, vol. 13, no. 1–2, pp. 119–163, 2019.

[12] D. Janßen and N. Langen, "The bunch of sustainability labels – do consumers differentiate?", *Journal of Cleaner Production*, vol. 143, pp. 1233–1245, 2017.

[13] A. Nygaard, "Is sustainable certification's ability to combat greenwashing trustworthy?", *Frontiers in Sustainability (Lausanne)*, vol. 4, 2023.

[14] A. Badhwar, S. Islam, C. Swee Lin Tan, T. Panwar, S. Wigley, and R. Nayak, "Unraveling green marketing and greenwashing: A systematic review in the context of the fashion and textiles industry", *Sustainability*, vol. 16, no. 7, 2024.

[15] M. Li and L. Zhao, "Exploring global fashion sustainability practices through dictionary-based text mining", *Clothing and Textiles Research Journal*, vol. 41, no. 3, pp. 175–190, 2021.

[16] L. Almeida, "Ecolabels and organic certification for textile products", in *Roadmap to Sustainable Textiles and Clothing*. Singapore: Springer, 2014, pp. 175–196.

[17] X. Lu, T. Sheng, X. Zhou, C. Shen, and B. Fang, "How does young consumers' greenwashing perception impact their green purchase intention in the fast fashion industry? an analysis from the perspective of perceived risk theory", *Sustainability*, vol. 14, no. 20, 2022.

[18] Changing Markets Foundation, *Licence to greenwash: How certification schemes and voluntary initiatives are fuelling fossil fashion, Report*, 2022. [Online]. Available: `https://changingmarkets.org/report/licence-to-greenwash-how-certification-schemes-and-voluntary-initiatives-are-fuelling-fossil-fashion/`.

[19] S. Alkhatib, P. Kecskés, and V. Keller, "Green marketing in the digital age: A systematic literature review", *Sustainability*, vol. 15, no. 16, 2023.

[20] Council of the European Union, *Empowering consumers for more sustainable choices*. [Online]. Available: `https://www.consilium.europa.eu/en/policies/green-claims-empowering-consumers-for-more-sustainable-choices/`.

[21] Y. Ni, "Design and implementation of an AI-based green marketing decision support system", in *Proceeding of the 2024 5th International Conference on*

Computer Science and Management Technology, New York, NY, USA: ACM, 2024, pp. 753–757.

[22] D. Moravcikova, A. Krizanova, and J. Kliestikova, "Green marketing as the source of the competitive advantage of the business", *Sustainability*, vol. 9, no. 12, 2017.

[23] J. McGuinn et al., *Environmental claims in the EU: inventory and reliability assessment: final report*. Luxembourg: Publications Office, 2024. [Online]. Available: `https://data.europa.eu/doi/10.2779/83089`.

[24] N. Nemes et al., "An integrated framework to assess greenwashing", *Sustainability*, vol. 14, no. 8, 2022.

[25] M. A. Delmas and V. C. Burbano, "The drivers of greenwashing", *California Management Review*, vol. 54, no. 1, pp. 64–87, 2011.

[26] Terrachoice, *The sins of greenwashing, Home and family edition*, 2010. [Online]. Available: `https://www.twosides.info/wp-content/uploads/2018/05/Terrachoice_The_Sins_of_Greenwashing_-_Home_and_Family_Edition_2010.pdf`.

[27] L. Alizadeh, M. C. Liscio, and P. Sospiro, "The phenomenon of greenwashing in the fashion industry: A conceptual framework", *Sustainable Chemistry and Pharmacy*, vol. 37, 2024.

[28] S. Fella and E. Bausa, "Green or greenwashed? examining consumers' ability to identify greenwashing", *Journal of Environmental Psychology*, vol. 95, 2024.

[29] Z. Yang, N. T. T. Huong, N. H. Nam, N. T. T. Nga, and C. T. Thanh, "Greenwashing behaviours: Causes, taxonomy and consequences based on a systematic literature review", *Journal of Business Economics and Management*, vol. 21, no. 5, pp. 1486–1507, 2020.

[30] Y. Bao, A. K. Obeid, D. Angus, J. Bagnara, and C. Leckie, "Shedding light on greenwashing: Explainable machine learning for green ad detection", in *AI 2024: Advances in Artificial Intelligence*, vol. 15442, Singapore: Springer, 2024, pp. 186–197.

[31] European Parliament, *Stopping greenwashing: How the EU regulates green claims*, 2024. [Online]. Available: `https://www.europarl.europa.eu/topics/en/article/20240111STO16722/stopping-greenwashing-how-the-eu-regulates-green-claims`.

[32] Agence de la Transition Ecologique, *Guide anti greenwashing*. [Online]. Available: `https://communication-responsable.ademe.fr/sites/default/files/2024-03/20230727_ademe_guide_antigreenwashing_web-vdef-min.pdf`.

[33] Agence de la Transition Ecologique: Le site de la communication responsable, *Communication rse : Évitez-vous les pièges du greenwashing ?* [Online]. Available: `https://communication-responsable.ademe.fr/test-communication-rse-greenwashing`.

[34] G. Marcatajo, "Green claims, green washing and consumer protection in the european union available", *Journal of Financial Crime*, vol. 30, no. 1, pp. 143–153, 2023.

[35] J. Kobti, V. Schmitt, and V. Woloszyn, "Towards automatic green claim detection", in *Conference: FIRE 2021: Forum for Information Retrieval Evaluation*, 2021, pp. 28–34.

[36] H. Kang and J. Kim, "Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods", *Applied Sciences*, vol. 12, no. 11, 2022.

[37] Y. Chen and M. Ding, "Detection of greenwashing in ESG reports of chinese listed companies based on word2vec and TF-IDF", in *Proceedings of the 2024 International Conference on Innovation in Artificial Intelligence*, New York, NY, USA: ACM, 2024, pp. 159–164.

[38] CSRD-directive, *Obligations imposed by the csrd directive*. [Online]. Available: `https://csrd-directive.eu/csrd-objectives-challenges/csrd-directive-obligations/`.

[39] European Comission, *Corporate sustainability reporting*. [Online]. Available: `https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en`.

[40] M. Mahdavi, R. B. Mehr, and T. Debus, "Combat greenwashing with goalspotter: Automatic sustainability objective detection in heterogeneous reports", in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, New York, NY, USA: ACM, 2024, pp. 4752–4759.

[41] S.-Y. Huang and Y.-M. Li, "Greenwashing detection mechanism based on multimodal analysis and social intelligence", in *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, IEEE, 2024, pp. 545–548.

[42] M. R. Maarif, M. Syafrudin, and N. L. Fitriyani, "Uncovering sustainability insights from amazon's eco-friendly product reviews for design optimization", *Sustainability*, vol. 16, no. 1, 2024.

[43] Eurostat, *Household budget survey - statistics on consumption expenditure*. [Online]. Available: `https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/80634.pdf`.

[44] T. Calamai, O. Balalau, T. Le Guenedal, and F. M. Suchanek, "Corporate greenwashing detection in text – a survey", *arXiv:2502.07541*, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.07541.

[45] S. Jäger, A. Flick, J. A. Sanchez Garcia, K. von den Driesch, F. Bießmann, and I. Trajanovska, *Greendb: A product-by-product sustainability database [data set]*, version 0.2.14, Zenodo, 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10607306.

[46] *Scikit learn API: Tfidfvectorizer*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

[47] S. Qaiser and R. Ali, "Text mining: Use of TF-IDF to examine the relevance of words to documents", *International Journal of Computer Applications*, vol. 181, no. 1, 2018.

[48] M. Andrés, H. Tejedor, and Á. Hernández, "Conditional and unconditional tests (and sample size) based on multiple comparisons for stratified 2×2 tables", *Computational and Mathematical Methods in Medicine*, pp. 1–8, 2015.

[49] M. Kateri, *Contingency Table Analysis, Methods and Implementation Using R*. Birkhäuser New York, NY, 2014.

[50] K. Rothman, K. Huybrechts, and E. Murray, *Epidemiology: An Introduction*. New York, NY: Oxford University Press, 2024.

[51] P. Webb, C. Bain, and A. Page, *The Mantel–Haenszel method for calculating pooled odds ratios*. Cambridge University Press, 2024, pp. 395–396.

[52] G. Tripepi, K. Jager, F. Dekker, and C. Zoccali, "Stratification for confounding –part 1: The mantel-haenszel formula", *Nephron. Clinical Practice*, vol. 116, no. 4, pp. c317–c321, 2010.

[53] Pennsylvania State University, Eberly College of Science, *STAT 504 / analysis of discrete data, Lesson 3: Two-way tables: Independence and association*. [Online]. Available: `https://online.stat.psu.edu/stat504/Lesson03`.

[54] G. Upton, *Categorical Data Analysis by Example*. Hoboken, New Jersey : Wiley, 2017.

[55] A. Ufondu, U. Shukla, C. Stambaugh, K. Huber, and N. Stambaugh, *Chapter 29 - Categorical variable analyses: Chi-square, Fisher's exact, Mantel–Haenszel*. Academic Press, 2023, pp. 165–170.

[56] A. Vierra, A. Razzaq, and A. Andreadis, *Chapter 28 - Categorical Variable Analyses: Chi-square, Fisher Exact, and Mantel–Haenszel*. Academic Press, 2023, pp. 171–175.

Use of AI

ChatGPT was used for formatting the LaTex-document, completing the list of generic sustainability claims, comparing different statistical models that are used together with odds ratios, and as a support for writing the Python code of the application.