Social Media Mining for Ideation: The Identification of Sustainable Solutions and Opinions

Sercan Ozcan [a, b, c], Metin Suloglu [d], C. Okan Sakar [e], Sushant Chatufale [f]

[a] Portsmouth Business School, University of Portsmouth, Portsmouth, UK

[b] Department of Engineering Management, Bahcesehir Universitesi, Istanbul, Turkey

[c] National Research University Higher School of Economics (Russian Federation)

[d] School of Computing, Leeds University, Leeds, UK

[e] Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey

[f] School of Engineering and Applied Science, Aston University, Birmingham, UK

e-mails: {sercan.ozcan@port.ac.uk; sc19ms@leeds.ac.uk; okan.sakar@eng.bau.edu.tr; s.chatufale@aston.ac.uk}

Corresponding Author: Sercan Ozcan, Portsmouth Business School, University of Portsmouth, Portsmouth, UK, e-mail: *sercan.ozcan@port.ac.uk*

Abstract

The availability of social media-based data creates opportunities to obtain information about consumers, trends, companies and technologies using text-mining techniques. However, the quality of the data is a significant concern for social media-based analyses. The aim of this study was to mine tweets (microblogs) to explore trends and retrieve ideas for various purposes such as product development, technology and sustainability-oriented considerations. The core methodological approach was to create a classification model to identify tweets that contained an idea. This classification model was used as a pre-processing step so the query results obtained from the application-programming interface were cleared from the messages that contained the search terms used in the query but did not contain an idea. The results of this study demonstrate that our method based on text mining, and supervised or semi-supervised classification methods, can extract ideas from social media. The social media data mining process illustrated in our study can be utilised as a decision-making tool to detect innovative ideas or solutions about a product or service and summarise them into meaningful clusters. We believe that our findings are significant for the sustainability, tech mining and innovation management communities.

Keywords: text mining; semi-supervised learning; support vector machines; decision-making; crowdsourcing; sustainability.

1. Introduction

Ideation is the most crucial and initial step of an innovation process [1]. It begins with the identification of an issue or the creation of a need for a product or service development process. The ideation process can occur within or outside of an organisation's endeavours. Nowadays, many firms utilise procedures and software to motivate and collect ideas internally [2]. This is an extremely common approach, particularly for large companies as they have access to immense human capital and resources that can be used for the creation of value in various phases, such as idea generation, product development or commercialisation [3]. Considering external resources, open innovation, co-creation and crowdsourcing are popular concepts and are utilised by companies to interact with consumers, inventors and other organisations to enhance their innovation capability [4–6].

A popular process in all open external approaches is to accumulate information from consumers for various stages of the innovation process. Many companies implement structured systems, methods and targeted competitions to obtain valuable input from consumers regarding their existing products or potential innovations. An innovative approach and a cost-effective alternative for collecting ideas from external resources is the utilisation of data obtained from social media platforms such as Twitter, Facebook and Instagram [7]. Social media is an excellent source for mining data, particularly if the required information concerns consumers and products.

The demand for product improvements, innovations and the scale of production is increasing to satisfy the needs of consumers and to aid companies in achieving their economic endeavours. Unfortunately, economic viability for many firms does not equal, or is not related to,

sustainable viability [8]. The economic orientation of firms has led to the growth of manufacturing and resource consumption resulting in increased negative effects on sustainability [9]. This is a critical issue and all relevant stakeholders must help to resolve the sustainability vs. economy game. Sustainability-related models have three primary pillars: economic, environmental and social sustainability [10] and circular economy models illustrate excellent methods of creating economically viable sustainable processes [11]. According to the literature on sustainability, it is best if there is earlier involvement of sustainability-related approaches in product or strategy development [12, 13]. Ideation is the initial stage of development of a new product or an innovation process. We believe that if ideation can be achieved via social media that focuses on sustainability, it will be possible to attain all three pillars of sustainability. The social pillar would be stronger due to the involvement of the social communities in the ideation process. The economic pillar would be achievable due to the indicated demand from the key stakeholders. Finally, there would be an expansive database from which environmentally sustainable ideas could be retrieved.

Considering the abundance of literature on sustainable approaches for product development [14–17], several studies have focused on the new product development (NPD) process through the lens of sustainable operations, manufacturing and processes. However, to date, no studies have focused on the sustainable ideation process and the method involved. Several studies have examined the ideation process from a variety of external data sources such as patents, publications and social media data [18–20]. However, few studies exist that focused on idea classification and identifying ideas for the NPD process. Pak and Paroubek [21] focused on opinion mining using a positive, negative and neutral classification system. Bifet and Frank [22] focused on knowledge discovery by classifying positive or negative feelings. Kruse et al. [23] proposed an idea mining approach that utilised an unsupervised clustering technique. Milosovic et al. [24] described a classification approach for social innovations using a naïve

Bayes classifier. Mirtalaie et al. [25] proposed a framework to generate innovative ideas based on the comments shared in social media with a cross-domain analysis. Similar to the present study, Christensen et al. [26] proposed the use of machine learning techniques to identify the user comments that contained an idea. In a recent study, Zhang et al. [27] suggested the use of deep learning techniques to detect sentences shared in online platforms that contained product-related innovative ideas. None of the above-mentioned studies proposed an end-to-end framework that would reveal clusters of ideas with a word network map including keywords related to product or service development.

Based on a gap in the literature and the above-mentioned limitation, this study aimed to mine Twitter data to explore trends and retrieve ideas for various purposes such as product development, technology and sustainability-oriented considerations. Twitter data from 2016 to 2018 containing different combinations of the hashtags #idea, #technology, #sustainability and #npd was compiled. The primary approach was to create a classification model to identify the tweets that contained an 'idea'. This classification model was used as a pre-processing step so that the query results returned by the Twitter application-programming interface (API) were cleared from the tweets that contained the search terms used in the query but did not contain an idea. For this purpose, we used various text processing and machine learning techniques. The study results demonstrate that our method, based on text mining and classification methods, can extract ideas from consumers. Moreover, it is an effective method for demonstrating technological trends and we believe our findings are significant for sustainability-related communities. Companies and entrepreneurs can utilise this method to identify information for product development activities that focus on sustainability endeavours.

The remainder of the paper is organised as follows: Section 2 provides a detailed literature review for the NPD and ideation processes, social media mining for NPD and sustainability-related NPD studies, Section 3 provides a detailed explanation of our social media mining

process and Section 4 provides sustainability-related ideation results. Finally, in Section 5, we conclude with key findings, implications, limitations and suggestions for future research.

2. Background and Literature Review

According to the European Commission's 2010 study on sustainable innovation, approximately 80–90% of environmental impacts are recognised during the design and development phases of the NPD process. This implies that NPD professionals should begin identifying sustainable problems or potential solutions from an early point using the appropriate procedures [28–30]. Therefore, procedures that enable the earlier involvement of sustainability approaches for NDP processes would be crucial for successful sustainable outputs. Accordingly, we first examined the NPD and ideation related process and literature. Next, we examined current investigations regarding sustainable NPD and ideation and, finally, we reviewed the literature in which NPD and innovation-oriented ideas were retrieved from external sources.

2.1 The NPD Process and Ideation

The NPD process concerns the management of the disciplines involved in the development of new products. The relevant models in this field describe activity regarding the development of a new product, service or solution that will be commercialised and, hence, there are certain variations, such as a service development process or a service innovation process. The NPD process involves various stakeholders of a company, such as product or innovation management, R&D management, marketing management and consumers who work together to continually advance the quality of products and innovation [31, 32].

New product development is a process through which an idea is transformed into a commercial output [33] and is, therefore, concerned with the renewal of products and services provided by organisations, and key determinants and how they are generated and delivered [34]. Currently,

studies on NPD are focused on commercialisation processes with shorter development cycles and high economic returns [35]. Typically, NPD activities include idea generation, market research, product design and engineering details [36]. However, a variety of NPD process models have been developed and introduced in the literature. Table 1 summarises several of these NPD process models, and the ideation phase can be seen as the first stage across all innovation and NPD processes.

Table 1. Summary of NPD process models

Wolf (1994) [37]	**Dimancescu and Dwenger (1996) [38]**	**Griffin (1997) [39]**	**Cooper (1998) [40]**	**Crawford and DiBenditto (2008) [33]**	**Trott (2017) [41]**
Idea conception	Idea	Idea/concept generation	Idea generation	Opportunity identification and selection	Idea generation
Awareness	Design	Idea screening	Preliminary assessment	Concept generation	Idea screening
Matching	Plan	Business analysis	Concept	Concept/project evaluation	Concept testing
Appraisal	Engineer	Development	Development	Development	Business analysis
Persuasion	Produce	Test and validation	Test trial	Launch	Product development
Adoption decision	Distribute	Commercialisation	Launch		Test marketing
Implementation	Dispose				Commercialisation
Confirmation					Monitoring and evaluation
Routinisation					
Infusion					

As illustrated in Table 1, the NPD process is based on a sequence of evaluative stages, and the key to a successful product development output involves the careful selection of a valuable idea that reduces the rate of failure at the end of the development of a product or process [42]. Initially, one may question what is accepted as an idea for an NPD process or for innovation. According to the Oxford Dictionary, an idea is a thought or suggestion as to a possible course of action. In the case of innovation, an idea is the source of novelty or development and it is the initiating point of the process [18]. In the present study, we accepted an idea as a potential solution or an opinion for innovation. We examined tweets that could lead to new opportunities for the development of innovations and products. We did not include comments by consumers

that did not include clear ideas. Idea sources include individuals or inventors, competitors, academic knowledge, feedback from consumers, internal R&D, innovation networks or clusters, open innovation and crowdsourcing approaches, suppliers and prior know-how [41]. Accordingly, social media is being considered as a new method for sourcing ideas.

To obtain the most benefit from the above-mentioned ideation sources, companies must increase the quantity (more ideas to be brought in) and/or quality (improved assessment methods and targeted involvement) of the ideation phase. An NPD process would end (or would not begin) if no opportunities were to arise from a pool of potential NPD or innovation ideas [43]. To reduce uncertainty in the ideation phase, ideas can be sourced from consumers who deliver their expectations for a new product or the next version of a product [43]. This is a key stage for resolving sustainability issues and obtaining sustainable innovations or NPDs. Su et al. [44] linked NPD with the consumer knowledge management process using a data mining approach. This study illustrated the significance of the requirements of specific consumer groups in the market for business excellence and product development success.

The rapid development and adoption of information technologies resulted in a new array of possibilities for organizations that are interested in adjusting activities outside and within the boundaries of the firm [45]. In the world of business and innovation discovery, ubiquitous connectivity has altered the nature of consumerism, allowing for the physical and non-physical parts of corporations to co-create, resulting in the creation of value for firms and their customers [46]. This notion is spurred by the consumer-centric culture of the internet which emphasises interactivity, speed, individuality and openness. Consumers can influence where and how value is generated and, in general, do not always see the value that a company believes a product is worth [46]. This sense of empowerment enables consumers to communicate through internet websites, e-mail and social networks thus providing ideas for new products or services that may improve on existing offerings in the market [47].

Research has shown that crowdsourcing is an interesting way to contribute to ideation, opportunity identification and concept development, due to an increase in the use of IT tools and platforms [48]. Observing and fulfilling the desires of consumers with new products has led to the creation of outstanding products and has shaped entire industries [49]. The concept of outsourcing has been infused with key methods that leverage the disruptive power of the crowd, such as the creative contribution of ideas and the identification of solutions to innovations (crowdsourcing) and the funding of innovative projects (crowdfunding) [45, 50]. As part of the crowdsourcing approach, the ideas or opinions of crowds are used for various purposes [51]. For example, Klein and Garcia [52] proposed an approach to improve the efficiency of the idea-filtering task performed on ideas shared in open innovation platforms. Dellerman et al. [53] proposed the use of machine learning methods to match the creative idea with the correct user. For this purpose, they combined the outputs of a machine learning model and a crowd evaluation to assign the ideas to humans. In a separate study, Banken et al. [54] proposed an automated idea allocation method that can be used as a decision support system by managers in innovation contests for optimal allocation of the ideas to raters.

All the relevant literature on crowdsourcing has focused on the process involving targeted ideation activities where competitions or challenges are used to encourage consumers to participate, mostly via established crowdsourcing platforms such as InnoCentive. However, the literature has also highlighted several issues such as the financial requirements and related motivational issues of consumers involved in the crowdsourcing process. Hence, we propose a social media mining model for ideation, where the required resources would be limited and the use of large-scale data with a targeted approach (i.e., classification methods) would help to minimise these issues.

2.2. Social Media Mining for NPD

The ideas or opinions of crowds retrieved from various open innovation or social media platforms have been used for a variety of purposes [51]. The evaluation and classification of innovation ideas in online platforms have been examined in several studies. Considering that the manual evaluation of ideas shared in online platforms is an extremely time-consuming and costly task, Dinh et al. [55] designed a framework that identifies potential ideas. First, they extracted the important terms from the idea texts retrieved from the crowdsourcing platform. Next, they evaluated the extracted terms using the numerical methods proposed in their study and, hence, converted the textual data to a set of features that could be inserted into a machine learning algorithm. In the final step, they used a mode logistic regression algorithm to compute the probability of the related idea having a potential for NPD.

In a recent study, Ma et al. [56] used the logistic regression algorithm to analyse the characteristics of the adopted and non-adopted user innovations shared in an online gaming community. The authors aimed to create a model to predict whether an innovation would be adopted. They provided a detailed analysis of the factors that influenced the adoption probability. Mirtalaie et al. [25] proposed a framework to generate innovative ideas based on the comments shared on social media. The authors performed a cross-domain analysis to explore the features that could be applied to the new versions of the analysed product. They used feature sentiment analysis tools and proposed new features based on the level of innovativeness and viability to assist the product developers in the product ideation phase.

Christensen et al. [26] also proposed a framework to identify user-generated content that contained an idea. The authors retrieved 3000 texts from an online community and two raters manually labelled the texts as ‘idea’ or ‘not idea’. Similar to the present study, there was a class imbalance issue since only a small number of texts (137) were labelled as ‘idea’. The authors combined text mining and machine learning approaches to create a model that could distinguish

whether a text contained an idea. They used the bag-of-words (BoW) technique for text representation with term frequency and binary term occurrence metrics. They also used support vector machines (SVMs) for the related binary classification task. To address the class imbalance issue, the authors did not create new samples with sampling techniques but instead used bootstrap aggregation to generate training sets with balanced class distribution.

In a separate study, Ko et al. [57] analysed social media-based consumer feedback to identify product opportunities by employing a topic modelling approach based on the latent Dirichlet allocation (LDA) technique. First, the authors applied topic modelling to determine the main topics related to the reference product. Next, they constructed a topic-based product graph using the co-occurrence data of the extracted topics and generated new product opportunities based on the obtained opportunity graphs. Jeong et al. [58] also used LDA for topic modelling, in addition to sentiment analysis and an opportunity algorithm to explore product opportunities. The satisfaction levels of the consumers were measured using a deep learning-based sentiment analysis approach. The proposed approach was applied to user reviews retrieved from social media that were related to a specific smartphone model.

In a recent study, Zhang et al. [27] suggested the use of deep learning techniques to detect sentences shared in online platforms that contain product-related innovative ideas. The authors developed an ensemble method that combined GloVe, XLNET and BERT to better represent the semantics and context of the related sentences. They inserted the obtained representations into a long short-term memory (LSTM) model to discriminate the innovative sentences. Similar to the present study, they also addressed the class imbalance issue because most user reviews did not contain an innovative idea. The authors integrated a focal loss function into the ensemble-based LSTM model to address this issue.

Considering the methods used in studies that utilised user-generated content in social media for NPD, several works utilised classification techniques to identify comments that contained

an idea. These studies were based on the construction of a fully supervised classification model in which all samples were used with their class labels during training. Considering that the labelling of social media data requires significant human effort, in contrast to these studies, we investigated the effectiveness of a semi-supervised learning approach. Moreover, we proposed a complete ideation framework for NPD, from data collection to word heatmap generation. Thus, we demonstrated how the identification of online content that does not contain an idea (using machine learning techniques) and the removal of this content from the retrieved data affects the quality of the resulting word heatmaps.

2.3 NPD, Ideation and Sustainability

Sustainable public policies are a key instrument in encouraging firms to move from economic to sustainability goals [9]. It is, therefore, imperative for organizations to integrate the sustainability concept into their NPD process. It is widely argued that integrating sustainability into business strategy or products can generate many benefits for the firm and society [9]. For example, an organization's corporate social responsibility programme, a key aspect of an organization's sustainability orientation, enables the firm to leverage its external stakeholder's knowledge to create the capability to generate new product innovations [59, 60]. Firms are encouraged by consumers who favour sustainable solutions, which ultimately results in operational efficiencies, improved product quality and greater customer value [61]. The success of NPD depends on the ability of manufacturing firms to create economic, environmental and social value for consumers and stakeholders [62] and this mostly occurs during the ideation and design stages. One method of achieving value is through sustainable innovation in the NPD process and, hence, via the generation of sustainable ideas. Pujari [63] observed that 'a higher degree of market focus' positively affected the market performance of eco-innovation activities. Hence, for particularly sustainable solutions, market assessments and the opinions of consumers are of greater importance for market success. Scientific solutions for sustainable

products are not adequate and the opinions of consumers must be considered regarding product features, design and performance.

The literature has shown that there is a lack of understanding concerning the sustainability aspects of NPD [61, 64]. Moreover, several studies concluded that there is a lack of understanding regarding the ability of firms to integrate the general sustainability concept into their NPD practices and at what stage of the processes the integration occurs [65, 66]. Aside from these issues, there is a methodological gap in the literature that enables sustainable ideation using external sources for NPDs. Most sustainable innovations focus on sustainable materials and manufacturing processes. However, incremental and radical sustainable innovations can be achieved by utilising the wisdom of the crowd and linking it to the earliest stage i.e., ideation.

2.4 Literature Gap and the Aim of the Study

In addition to an extensive literature review on NPD, ideation and relevant sustainability studies provided in Sections 2.1 and 2.2, we also reviewed the methodological approaches where social media was mined for ideation. Accordingly, several studies attempted to aid in ideation or accumulate opinions using social media data [21-27]. However, the present study differs from these works through the development of an end-to-end framework that can aid decision-makers in retrieving and identifying NPD ideas from a word network map including clusters of keywords. Moreover, none of the above-mentioned studies considered sustainability.

Considering the literature review presented in Sections 2.1 and 2.2 and the above-mentioned methodological and practical gaps, this study aimed to establish the first social media mining process for ideation that focuses on sustainability ideas and opinions. The conceptual framework of the study is presented in Figure 1. The objectives of the study were as follows:

- To establish a classifier model to mine ideas from social media,
- To examine the ideas and opinions of the sustainability community,
- To cluster and illustrate potential sustainable innovations.

Fig 1. Conceptual framework of the study

3. Materials and Methods

3.1. Dataset

The dataset consisted of 22891 tweets, collected using the Twitter API by querying tweets containing the words 'idea' and 'sustainability' and the hashtags #idea and #sustainability. The following search query was used to retrieve the relevant data:

Search query: (Idea OR #idea) AND (sustainabil* OR #sustainabl*)

1199 of the tweets were manually labelled as positive class (idea) or negative class (not an idea), leaving 21692 unlabelled tweets. The number of negative class tweets (927) was 3.4 times greater than positive class tweets (272) in the labelled portion of the dataset.

3.2. Research Method

The research methodology followed in this study is presented in Figure 2. The steps of the proposed ideation system, which are detailed in the following subsections, can be summarised as:

- The collection of 22891 relevant tweets using the Twitter API.

- Text pre-processing operations such as the removal of webpage links, e-mail addresses and other unnecessary symbols (such as punctuation). Exclamation marks and question marks were not removed.
- Tokenization, stop word removal and stemming operations using the Porter algorithm.
- Representing tweets using a classical text representation technique, the term frequency-inverse document frequency (TF-IDF) statistic, and BERT (a state-of-the-art word embedding method).
- Manual labelling of 1199 tweets as positive or negative class.
- Oversampling using the synthetic minority oversampling technique (SMOTE) on the labelled dataset to balance the number of positive and negative class samples.
- Training supervised and semi-supervised algorithms on the resampled dataset and obtaining classification models.
- Applying the classification models to the remaining unlabelled tweets and creating a new set that includes the tweets that are labelled as positive.
- Creating heatmap visualisations using all the tweets that are labelled as positive.

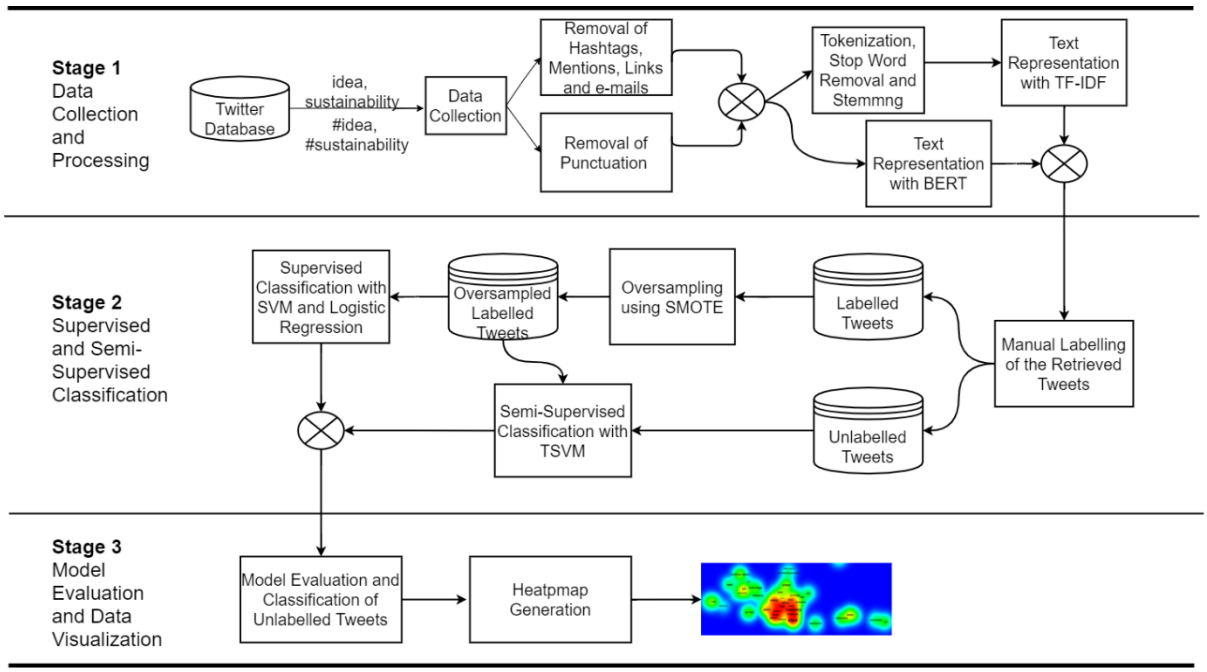

Fig 2. Overview of the proposed ideation system

3.2.1. Text Pre-processing and Representation

Before giving the tweets to machine learning classifiers, various natural language processing operations were applied to the dataset in the first stage of the proposed ideation system as seen in Figure 2. First, we removed all hyperlinks, e-mail addresses, user tags and hashtags as these do not contain important information for the related task. We also eliminated all punctuation characters from the tweets, however, the exclamation mark (!) and question mark (?) symbols were not removed due to the belief that these characters potentially provided additional predictive power to the classifiers. Furthermore, all characters in the corpus of the tweets were converted to lowercase and all stop words were removed. Porter's Snowball stemmer [67] was applied to all words after tokenization.

We utilised two text representation techniques. First, we applied the BoW technique, which involves representing a text or document as a series of words by disregarding the order information [68]. Various numerical statistics such as term frequency, TF-IDF and term variance can be used to represent the weight of each word in a document [69]. In the present

study, we used a BoW representation of the tokenized tweet text using TF-IDF. The TF-IDF statistic is a common representation technique used to determine to what extent a word is important in a document. A key issue with TF-IDF is that it ignores the context in which a word appears and the semantics of that word. The set of contiguous sequences of n words in a document, known as 'n-grams', can also be used to obtain a representation while considering the context in which a word appears. However, the disadvantage of this is an increase in the dimensionality of the feature space. We used scikit-learn API for Python [70] to apply the TF-IDF feature extraction technique on our corpus using unigram and bigram representations of all words to obtain a sparse vector representation for each tweet.

In addition to the classical TF-IDF method, we also used a state-of-the-art language representation model known as BERT [71] for text representation which was created using transformer architecture. The transformer encoder is a popular attention model trained bidirectionally in BERT to provide a better representation of the context of the language. Instead of predicting only the next word in the given sequence, BERT uses a masking mechanism. Through this approach, 15% of the input words are masked and the network is trained to predict the masked words from the non-masked words. The other important mechanism utilised in BERT is the next sentence prediction task in which pairs of sentences are fed to the network as input, and the network is trained to predict whether the given sentences followed each other in the original text. Positive and negative samples are created by providing sentence pairs that are next to each other and not next to each other. Due to the working mechanism of BERT, as seen in Figure 2, the pre-processing operations applied to the Twitter data before BERT were different from that of the TF-IDF representation. In this study, we used BERT-base (as a pre-trained model consisting of six encoders) to obtain the word representations in the given tweets. The BERT model was applied to the entire dataset including the labelled and unlabelled tweets.

3.2.2. Classification with Supervised and Semi-Supervised Learning

As described in Section 3.1, the dataset used in the present study consisted of 22891 tweets. We manually labelled 1199 of the tweets as positive class (idea) or negative class (not an idea). Since manual labelling is an extremely time-consuming task, in the second stage of the ideation system shown in Figure 2, we trained supervised and semi-supervised classifiers on the labelled dataset and applied the obtained models on the remaining 21692 unlabelled tweets to obtain their labels. The tweets that did not contain an idea were subsequently removed and only the tweets containing an idea were visualised via a word network map for ideation to be used in the NPD process.

Classification in a supervised learning setting utilises labelled samples to generate a hyperplane that separates the two classes. Many techniques can be used to generate this hyperplane, such as logistic regression, SVMs, artificial neural networks, and decision tree-based methods such as random forest. These specific techniques and many other machine learning classifiers aim to minimise a certain loss function which is dependent on the weights of the model. We used two fully supervised learning algorithms, an SVM and logistic regression, for the classification task in our experiments.

For larger numbers of samples, supervised methods are effective at modelling a variety of classification problems. However, obtaining the labels for each sample is not always straightforward. The lack of a sufficient number of labelled samples may result in an overfitted discriminating hyperplane. Due to this, semi-supervised methods that utilise a vast number of unlabelled samples during training have gained popularity in recent years [72]. In the present study, we used a semi-supervised learning algorithm for the classification task and compared it to the supervised learning algorithms. Thus, we aimed to utilise the unlabelled tweets during training to obtain a more generalizable classification model.

An SVM classifier is a machine learning algorithm that attempts to obtain a hyperplane that discriminates the samples belonging to positive and negative classes. The aim is to obtain a separating hyperplane such that the distance of the closest class points to this hyperplane will be maximised. This distance is known as the margin. The samples, which can be considered the most difficult to classify, are known as support vectors [73]. By employing a method known as the 'kernel trick', the inputs can be mapped to new feature spaces and non-linear classification can also be performed. Given n samples $x_i \in \mathbb{R}^d, i = 1, \dots, n$ with two distinct classes and their corresponding labels $y_i \in \{-1, 1\}, i = 1, \dots, n$, the constrained optimization problem SVMs solve can be written as:

$$min_{\boldsymbol{w},b,\xi} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n}\xi_i$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, i = 1, \dots, n \qquad (1)$$

where C is a regularization parameter of the SVM that is fine-tuned to control the complexity of the model, ϕ is the kernel function and ξ_i are the slack variables introduced to tolerate misclassifications in the training set. Classification tasks with more than two labels can be divided into multiple binary classification problems. Other methods of performing multi-class classification have also been proposed [74, 75].In the present study, we also used an SVM-based semi-supervised learning approach known as Transductive SVM (TSVM). While traditional SVMs only process labelled data during training, TSVMs can utilise unlabelled data points and aim to move the classification hyperplane towards low-density regions in the feature space. Thus, a more generalizable model can be obtained, particularly where there is a limited number of labelled samples and a large number of unlabelled samples. The typical constrained optimization function for a linear TSVM is as follows:

$$min_{\boldsymbol{w},b,\xi,\xi^*,y^*} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi_i + C^*\sum_{i=l+1}^{l+u}\xi_i^*$$

subject to $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, 1 \leq i \leq l$

$y_i^*(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i^*, l+1 \leq i \leq l+u$

$\xi_i \geq 0, 1 \leq i \leq l$

$\xi_i^* \geq 0, l+1 \leq i \leq l+u$ (2)

where l and u are the number of labelled and unlabelled samples, respectively.

A TSVM algorithm for text classification was proposed by Joachims [76]. During training, the algorithm first assigns temporary labels to the unlabelled samples using an inductive SVM and then identifies the unlabelled samples with a positive and negative label such that switching the labels of these two samples would result in a decrease in the optimization function output. Sindhwani and Keerthi [77] began with this algorithm and proposed the use of the modified finite Newton l_2-SVM method [78] during optimization for faster execution. Furthermore, instead of switching the labels of the unlabelled samples one at a time, they proposed a multiple switching approach to switch the labels of multiple unlabelled samples concurrently during training. In our experiments, we used the linear multi-switch TSVM described by Sindhwani and Keerthi [77]. This approach was specifically designed with large sparse datasets in mind. Considering that BoW representations of tweets are highly sparse matrices (since tweets consist of short texts containing a limited number of characters), particularly for TF-IDF representation, the TSVM approach was an appropriate fit for our dataset.

3.2.3. Oversampling using SMOTE

As mentioned in Section 3.1 of our study, there was an imbalance between the number of positive and negative class tweets in the labelled portion of our dataset, as the fraction of labelled tweets that involved ideas was less than those that did not involve ideas. There was approximately 3.4 times more negative class tweets (927) than positive class tweets (272). It is known that uneven class labels negatively affect the performance of machine learning models due to the growing tendency of the models to predict the majority class as the difference

between the number of samples of each class increases [79]. This issue is often referred to as the 'class imbalance problem'. To overcome this issue, several oversampling and undersampling techniques have been proposed. In our framework, we applied an oversampling technique known as SMOTE [80] to the labelled dataset to generate artificial positive class tweets that were similar, but not identical, to the real positive samples. The SMOTE algorithm used in our study is presented as Algorithm 1. SMOTE creates new samples by interpolation in the neighbourhood of the minority class instances. Specifically, as seen in Algorithm 1, a new point is generated on the vector between the original point and one of the randomly selected nearest neighbours. The exact position of the new point on this vector is determined using the λ parameter. The most common value for the size of the neighbourhood, k, is 5 [81], which was also utilised in our experiments. We set N to 3 to triple the minority class samples and obtain a balanced class distribution. We used a Euclidean distance metric to obtain the nearest neighbours of a given sample.

Algorithm 1

Input
x: set of minority class samples
N: number of artificial samples generated for each original sample
n: number of attributes
k: Number of nearest neighbours
Output
S: set of synthetic samples
$S \leftarrow \emptyset$
for each sample x_t at x
  Find k nearest neighbours of the sample
  for i from 1 to N
    Randomly choose one of the k nearest neighbours of x_t: $\bar{x}_t$
    for j from 1 to n
      $d = \bar{x}_t[j] - x_t[j]$
      Generate λ randomly from the range $[0, 1]$
      $x_{new}[j] = x_t[j] + \lambda * d$
    end for
    Insert generated artificial sample, x_{new}, to S
  end for
end for

3.2.4. Evaluation Metrics

As stated previously in Section 3.2.3, SMOTE is applied to balance the distribution of class labels in the training set. An important point to note is that this sampling procedure was not performed on the cross-validation or test sets to avoid misleading results during model evaluation by considering new artificial synthetic instances.

While accuracy is a useful and widely used metric to evaluate the quality of a prediction obtained via a machine learning model, the class imbalance issue impedes the expressiveness of this measure. Due to this, other metrics such as the F1 score have been suggested for obtaining a deeper understanding of the predictive capability of trained machine learning models. In addition to accuracy, we used the F1 score which conveyed the balance between the precision and recall measures.

3.2.5. Visualisation and Interpretation

We applied the heatmap visualisation method separately to two different sets of tweets. The first set included all 22891 retrieved tweets without any eliminations. The second set included only the tweets that were labelled as an idea by the best classifier obtained in our experiments. The cluster formation was based on the co-occurrence measurements of the terms that appeared in the set of tweets. First, a co-occurrence matrix of terms was calculated for the terms that had a minimum term frequency of two. A co-occurrence matrix represents the number of times one term appears with the other terms across documents. The results of the co-occurrence matrix calculations were used to calculate the centrality measurements [82, 83] using UCINET software. Centrality measurements position the terms based on their relevance to each other and their overall relationship to other terms. The relevance of terms leads to the closeness of nodes (terms) to each other. If one term has a low co-occurrence across tweets with another term, then these terms are positioned apart from each other.

To further explain the centrality measurements, the degree centrality calculates the number of links between terms, the closeness centrality calculates the average length of the shortest path between the nodes and all other nodes in the graph, and the betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Accordingly, all terms are positioned based on the centrality measurements, and the node sizes are adjusted based on the term frequency (how many times it appears). The positioning of terms based on the centrality calculations leads to a cluster of terms. Centrality measurements are used with VOSviewer software for visualisation using the clustering method. VOSviewer software also helps to improve the visualisation results by adjusting the size of the terms and allocates a colour code for each cluster based on the centrality measures.

As part of our interpretation step, we used single terms or a combination of terms to locate relevant tweets in the database. For example, the visualisation step showed the term 'farm' and using this term, many agriculture-related sustainability ideas were identified in the selected database. Some of the tweets had the relevant hashtags and so they may not have appeared in the actual idea statement. The interpretation step was completed by identifying illustrative ideas from the database using the terms that appeared in the clustering visual.

The reliability and the validity of the applied approach were ensured through consideration of the following points: 1) a large dataset covered a variety of thoughts of the relevant stakeholders in social media, 2) the labelling process was completed separately for three individuals for the same training set and disagreements in the labels were completed following a consensus-based decision-making process, 3) once all the tweets were labelled using a semi-supervised approach, approximately 10% of the positive tweets (ideas) were examined and the labels were satisfactory with minimal errors, 4) the noise (irrelevant results) in the final visuals was minimised due to a frequency-based approach and minimal thresholds and 5) using the

terms or a combination of terms, the results were crosschecked with the database for interpretation and to confirm that the results were reliable and valid for the purpose of the study.

4. Results

In this section, we first provide the results of the supervised and semi-supervised classification models used to explore the messages that contained an idea. Next, we present an analysis of the sustainability-related ideas that were grouped under different clusters using the heatmap visualisation method.

4.1. Classification

4.1.1. Experimental Setup

As mentioned in Section 3.1, there were 1199 labelled tweets in our dataset. However, considering that the most time-consuming part of such applications is the labelling process, we performed the classification experiments for various sizes of training sets. Thus, we aimed to demonstrate how supervised and semi-supervised algorithms performed with varying numbers of training examples. Moreover, we used a large portion of the dataset (900 of 1200 tweets) for training to obtain the best possible classification model and presented the heat maps generated using all the tweets, including the unlabelled ones that were labelled by the most effective model.

For cross-validation, we first divided the dataset consisting of the labelled tweets into two parts: the training set and the test set. Next, we applied SMOTE to the training set to obtain a balanced distribution between the positive and negative class samples. For validation, we applied a grid-search procedure on the training set to obtain the optimal values of the hyperparameters using a 5-fold cross-validation. Finally, the model trained with the best values of the hyperparameters was applied to the test set. This procedure was repeated 10 times for statistical significance and the average results obtained for the test set are presented in the following subsection.

4.1.2. Results

Figure 3 shows the average results of the semi-supervised and supervised linear classifiers for increasing numbers of training samples. Since SMOTE was applied only to the training set to produce over-optimistic results for the artificially created samples, the test set remained imbalanced. Therefore, while assessing the performance of the classifiers, the F1 score, which was a suitable metric for the imbalanced dataset (as it accounts for the precision and the recall) was considered to be of greater importance than accuracy.

As seen in Figure 3, in overall TSVM with BERT has given the highest F1 score and accuracy values. It is also seen that both supervised and semi-supervised classifiers gave significantly higher F1 scores with BERT representation compared to TF-IDF representation showing that BERT is more successful in representing the contextual information in the dataset. The highest F1 score (0.56) and accuracy (78%) were obtained using 400 samples for training with the TSVM's BERT-based semi-supervised model.

Another important metric that should be considered in our problem is precision since we use the classification algorithm with the aim of cleaning the Twitter API search results by eliminating the tweets that do not include an idea and only focusing on the tweets that contain innovative ideas. A model with a high precision will have a lower false positive value and hence a clearer word network including terms mostly from innovative ideas. The results in Figure 3 show that TSVM with BERT features has achieved a higher precision than the other algorithms especially when the number of training samples is less than 300. These results verify that semi-supervised approach can be preferred over supervised for the given task.

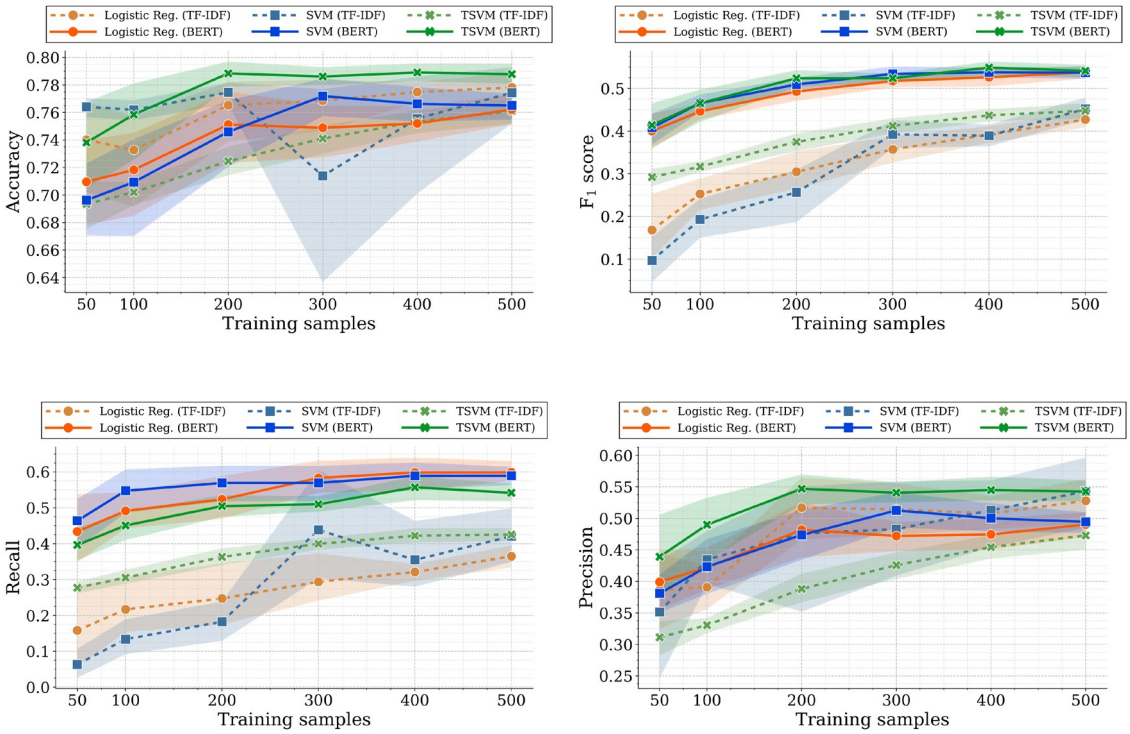

Fig 3. Performance of the classifiers with SMOTE with respect to the number of training samples

We repeated the same cross-validation procedure 10 times using 900 samples for training and the remaining samples for testing. Through this analysis, we aimed to construct the best possible model with the labelled dataset we had, using a large part of the dataset for training. These experiments were conducted with BERT representation only since it was shown to perform better than TF-IDF in the previous experiments. Table 2 shows the average and standard deviations of the evaluation metrics obtained using 900 training samples. The highest F1 score (0.564) was achieved via a supervised SVM. However, the differences between the F1 scores were not statistically significant. The results also demonstrate that the highest accuracy (78.4%) and precision (0.526) were obtained via the TSVM model. The recall of the logistic regression was higher than the other two classifiers. Additionally, the logistic regression had the lowest standard deviation for the F1 score which demonstrates that this algorithm was more robust for changing sets of training samples due to its simplicity.

Table 2. Performance of the classifiers using 900 training samples and BERT for text representation

		Accuracy	Precision	Recall	F1 Score
Logistic Regression	**Average**	0.761	0.486	**0.655**	0.557
	Standard Dev.	0.025	0.042	0.020	0.026
Support Vector Machine	**Average**	0.773	0.505	0.641	**0.564**
	Standard Dev.	0.025	0.042	0.048	0.038
Transductive SVM	**Average**	**0.784**	**0.526**	0.596	0.558
	Standard Dev.	0.027	0.053	0.053	0.049

4.2. Illustration of Ideas

This results section illustrates the ideas obtained from social media using a heatmap visualisation method. We utilised the BERT+TSVM model for this purpose due to its higher accuracy and precision compared with the two other classifiers (as shown in Table 2). We applied the most effective TSVM model to the 21692 unlabelled tweets and identified the tweets that contained an idea. There are two visualisations. Figure 4 shows the results for the unlabelled tweets and depicts a word map for all the tweets retrieved in our study. Figure 5 shows a word map based solely on the tweets that were classified as positive by the TSVM model. The results obtained via the TSVM demonstrate that it was a much cleaner and organised representation of the ideas. The results also show that using a classifier as a pre-processing method to clean the dataset (via the removal of tweets that did not contain an idea) resulted in an improved word network map for the ideation task.

Based on Figure 5, the results can be grouped into four main categories: 1) Eco-friendly and sustainable production and farming, 2) Sustainability and climate change education, 3) Sustainable packaging and materials and 4) Sustainable energy and building designs. Figure 5 also demonstrates the interrelationship between these four main clusters. For example, cluster 1 is close to cluster 2 as there is a strong connection between sustainable food production

(cluster 1) and educating children about a sustainable environment, sustainable gardening, and the environment and climate change (cluster 2). Clusters 1 and 3 are adjacent to each other due to environmental sustainability ideas in cluster 1 having sustainable production and cluster 2 having sustainable and recyclable packaging. Clusters 2 and 4 are adjacent to each other due to the relevancy of the ideas concerning sustainability and climate change education (cluster 2), and the use of environmental, sustainable and renewable energies (cluster 4).

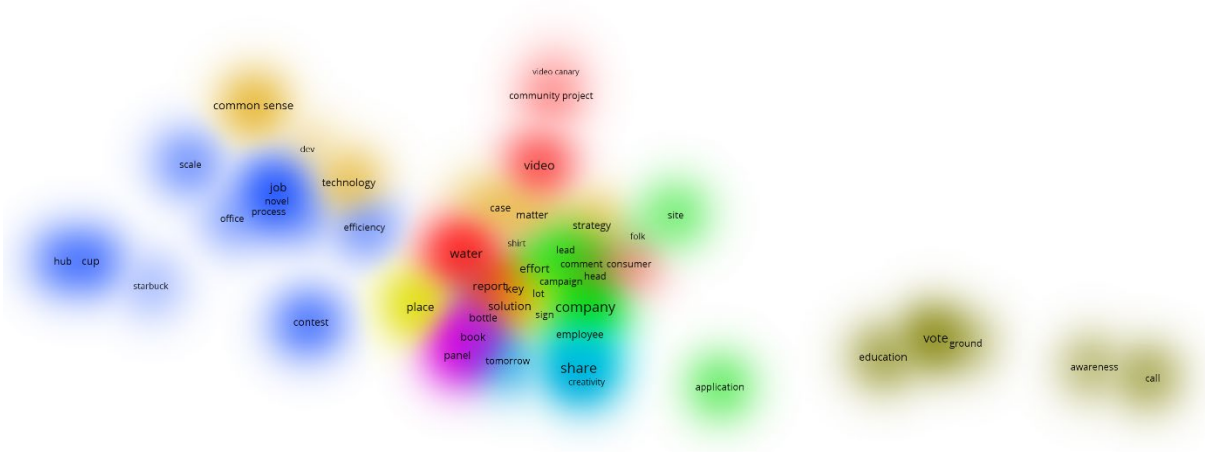

Fig 4. Visualisation of ideas without classification

Table 3 lists examples of ideas that were compiled for these four main categories. Cluster 1 contains many examples concerning sustainable farming models, particularly for urban locations as it is becoming more difficult to produce fruits or vegetables in an area where farming land is limited. For example, vertical farms, rooftop fish and greenhouse models for cities can be observed in cluster 1. Cluster 2 illustrates many examples concerning sustainability education and it is mostly focused on early age schoolchildren. There is a variety of sustainability education examples such as sustainable food consumption, electric vehicles and solar panels. Summer camps that involve sustainability and solar education weeks represent interesting education models.

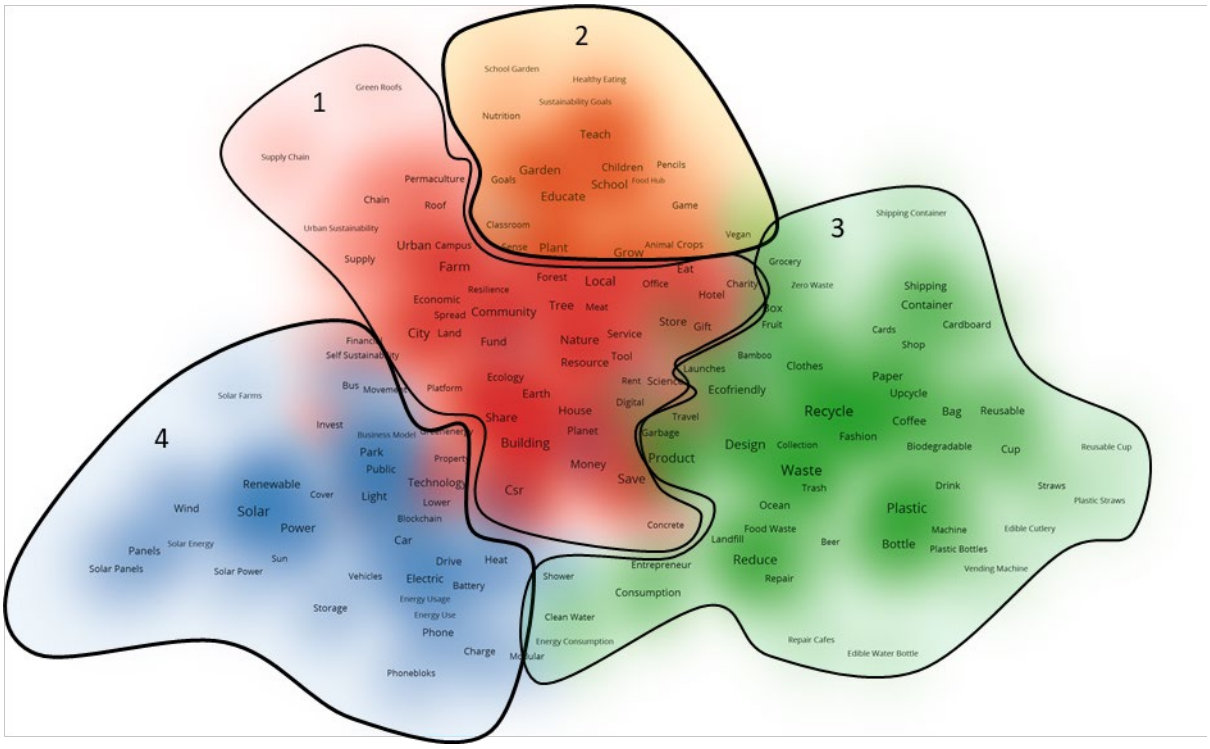

Fig 5. Visualisation of ideas after SVM-based classification

Cluster 3 contains many packaging solutions including three groupings of sustainable packaging which are: 1) reusable packaging, 2) recyclable packaging and 3) edible packaging. Of these three solutions, edible packaging materials were the most recently developed solution. Several packaging solutions targeted sustainability, not only by including a sustainable material but also by focusing on sustainable outputs. For example, supermarket vegetable packaging that can be used to 'grow your own' vegetables. Furthermore, numerous ideas for sustainable cups were shared on social media due to an online competition created by Starbucks. Cluster 4 consists of ideas concerning sustainable energy solutions. Many of the ideas obtained targeted the conditions and areas where electricity can be produced but is not being used as a renewable energy source. Several of the ideas involved the production of electricity via exercise bikes, dance floors and busy walking areas. Numerous sustainable energy production ideas focused on abandoned areas or frugal innovation approaches such as enabling underdeveloped countries to produce their own electricity.

Table 3. Illustration of sustainability ideas

Categories of ideas	Examples
Cluster 1 - Sustainable production	• An online message board for sustainable farmers • Meat idea: tracking produce back to farmers • Urban sustainability using vertical farms • Rooftop fish: the future of urban farming • The use of greenhouses in cities • Window farms • 'Farm from a box' – a solar-powered farm using a modified shipping container • Leather from pineapples • A solar-powered floating farm • Pollinator-friendly solar sites • Vertical ocean farming and fishing
Cluster 2 - Sustainability education	• Education regarding sustainability and healthy eating habits • Early education of children regarding sustainability • Solar education week • Electrical vehicle education • A summer camp that encourages sustainability education • Education of children concerning energy efficiency and sustainability
Cluster 3 - Sustainable packaging	• Karma cup system from Starbucks • Cups into compost for urban gardens • Edible water bottle • Edible food packaging • A convertible pizza box with built-in plates • Supermarket vegetable packaging that allows one to grow their own vegetables • Packaging based on mushrooms
Cluster 4 - Sustainable energy	• Dance floors that generate electricity • Exercise bikes that create energy for a sustainable gym • Helping developing countries to build wind turbines from scrap car parts • Using abandoned spaces to promote sustainable energy solutions • Glowing bio-LED trees to replace streetlamps • Environmentally friendly Lego-type bricks for the construction of houses

In summary, many sustainability-related ideas were observed; several of these were the ideas of individuals and some were the ideas of individuals who were promoting sustainability. Both

sources of ideas can aid communities in discovering potential routes for increasing sustainability at the individual, organisational and national level. The results of this study demonstrate that the sustainability community is pursuing ways to be more sustainable, particularly in urban areas where farming can be implemented. Moreover, many individuals are pursuing sustainable solutions for environmentally unsustainable materials due to the frequent use and disposal of these items. Finally, our results show that organisations that create online challenges to encourage sustainability accumulate a larger number of more valuable ideas.

5. Conclusions

The results presented in this study on social media mining for sustainability ideas can aid in resolving the misconception that data obtained from social media are of low quality and are limited to providing directions for organisations and relevant communities. To our knowledge, this is the first study to illustrate how ideas can be retrieved using social media mining for integration into product or packaging innovations. This study also demonstrates how the sustainability of communities can be enhanced by considering early education and the use of renewable energy at various conditions and locations.

This study provides methodological and practical contributions. Its practical contributions apply to those who are involved in innovation management, product development and sustainability. Its methodological contributions apply to tech mining, scientometrics and social media data analysis-related communities. Considering the practical contributions, our study highlights many examples for sustainability-related ideas, and these ideas can be used for product, service or business solutions to achieve global sustainable goals. The examples presented in clusters 1 and 3 (Figure 5) can be used for sustainable production and packaging solutions. The examples presented in cluster 2 (Figure 5) can be used for sustainability

adaptations. The examples shown in cluster 4 (Figure 5) can be used to develop innovative methods for the production and use of sustainable energy.

Considering the methodological contributions, we successfully created a classification model to identify the tweets that contained ‘an idea’ and ‘not an idea’. This classification model was used as a pre-processing step so that the query results returned by the Twitter API were cleared from the tweets that contained the search terms used in the query but did not contain an idea. To our knowledge, this is the first social media mining model developed to retrieve ideas, and this is one of very few studies in which a classification model was used as part of a pre-processing step.

Our study also provides implications for companies. The results indicate areas in key communities where there is a demand for innovations. As sustainable progress is challenged due to contradictions between sustainable and economic goals, these ideas may provide opportunities for organisations to prioritise areas where they can benefit from all three pillars of sustainability. Moreover, the results suggest excellent directions for the education sector, with potential topics such as early education and community-targeted sustainability education.

The limitations of our study concern the methodological process. We utilised classification models with 900 labels to identify the ideas that were specific to the field of sustainability, therefore, this process would have to be adjusted to retrieve ideas from other fields, if required. Accordingly, future studies may employ a similar method to identify ideas in other fields such as the electronics industry. Moreover, future studies could contribute to the field by creating suitable hashtags and classification methods for other ideation fields. In this study, we used the idea retrieval process as a supportive mechanism by utilising social media data for innovation and NPD activities. However, other relevant approaches, such as focus groups, brainstorming and Delphi methods with experts can also result in new ideas. Additionally, other studies that

utilise the idea retrieval approach employed in the present study can obtain sector- or product-specific sustainability ideas.

Acknowledgement

The article was prepared within the framework of the Basic Research Program of the National Research University Higher School of Economics.

References

[1] J. van den Ende, L. Frederiksen and A. Prencipe, "The front end of innovation: Organizing search for ideas," Journal of Product Innovation Management, vol. 32, no. 4, pp. 482–487, 2015.

[2] P. M. Di Gangi, and M. Wasko, "Steal my idea! Organizational adoption of user innovations from a user innovation community: A case study of Dell IdeaStorm," Decision support systems, vol. 48, no. 1, pp.303-312.

[3] A. Urbinati, M. Bogers, V. Chiesa, and F. Frattini, "Creating and capturing value from Big Data: A multiple-case study analysis of provider companies," Technovation, vol. 84, pp. 21-36, 2019.

[4] E. Loukis, Y. Charalabidis and A. Androutsopoulou, "Promoting open innovation in the public sector through social media monitoring", Government information quarterly, vol. 34, no. 1, pp. 99-109, 2017.

[5] K. Robson, M. Wilson, and L. Pitt, "Creating new products from old ones: Consumer motivations for innovating autonomously from firms", Technovation, vol. 88, p.102075, 2019.

[6] S. Kamboj, B. Sarmah, S. Gupta, and Y. Dwivedi, "Examining branding co-creation in brand communities on social media: Applying the paradigm of Stimulus-Organism-Response", International Journal of Information Management, vol. 39, pp.169-185, 2018.

[7] B. Krishnamurthy, P. Gill, M. Arlitt, "A few chirps about twitter," in Proceedings of the first workshop on Online social networks, pp. 19–24, 2008.

[8] S. Ghazinoory, A. Sarkissian, M. Farhanchi and F. Saghafi, "Renewing a dysfunctional innovation ecosystem: The case of the Lalejin ceramics and pottery", Technovation, p.102122, 2020.

[9] R. M. Dangelico, P. Pontrandolfo, and D. Pujari, “Developing Sustainable New Products in the Textile and Upholstered Furniture Industries: Role of External Integrative Capabilities,” Journal of Product Innovation Management, vol. 30, no. 4, pp. 642–658, 2013.

[10] M. A. Kahn, Concepts, definitions, and key issues in sustainable development: the outlook for the future, in Proceedings of the 1995 International Sustainable Development Research Conference, 1995.

[11] N. M. Bocken, I. de Pauw, C. Bakker, and B. van der Grinten, “Product design and business model strategies for a circular economy,” Journal of Industrial and Production Engineering, vol. 33, no. 5, pp. 308–320, 2016.

[12] P. Bansal, “Evolving sustainably: A longitudinal study of corporate sustainable development,” Strategic Management Journal, vol. 26, no. 3, pp. 197–218, 2005.

[13] M. Doelle and A. J. Sinclair, “Time for a new approach to public participation in ea: Promoting cooperation and consensus for sustainability,” Environmental Impact Assessment Review, vol. 26, no. 2, pp. 185–205, 2006.

[14] D. Maxwell and R. van der Vorst, “Developing sustainable products and services,” Journal of Cleaner Production, vol. 11, no. 8, pp. 883–895, 2003.

[15] P. R. Kleindorfer, K. Singhal, and L. N. Van Wassenhove, “Sustainable operations management,” Production and Operations Management, vol. 14, no. 4, pp. 482–492, 2005.

[16] A. Jayal, F. Badurdeen, O. Dillon Jr, and I. Jawahir, “Sustainable manufacturing: Modeling and optimization challenges at the product, process and system levels,” CIRP Journal of Manufacturing Science and Technology, vol. 2, no. 3, pp. 144–152, 2010.

[17] N. P. Melville, “Information systems innovation for environmental sustainability,” MIS Quarterly, vol. 34, no. 1, pp. 1–21, 2010.

[18] R. G. Cooper and S. Edgett, “Ideation for product innovation: What are the best methods,” PDMA Visions Magazine, vol. 1, no. 1, pp. 12–17, 2008.

[19] Y. Wich, J. Warschat, D. Spath, A. Ardilio, K. König-Urban and E. Uhlmann, "Using a text mining tool for patent analyses: Development of a new method for the repairing of gas turbines," 2013 Proceedings of PICMET '13: Technology Management in the IT-Driven Services (PICMET), pp. 1010-1016, July 2013.

[20] C. Moser, J. M. Birkholz, D. Deichmann, I. Hellsten, and S. Wang, “Exploring ideation: Knowledge development in science through the lens of semantic and social networks,” in 46th Hawaii International Conference on System Sciences, pp. 235–243, IEEE, 2013.

[21] A. Pak, P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." In LREc, vol. 10, no. 2010, pp. 1320-1326. 2010.

[22] A. Bifet and E. Frank, “Sentiment Knowledge Discovery in Twitter Streaming Data,” Discovery Science Lecture Notes in Computer Science, pp. 1–15, 2010.

[23] P. Kruse, A. Schieber, A. Hilbert, and E. Schoop. "Idea mining–text mining supported knowledge management for innovation purposes.", Proceedings of the Nineteenth Americas Conference on Information Systems, pp. 1-10, 2013.

[24] N. Milosevic, A. Gok, and G. Nenadic, “Classification of Intangible Social Innovation Concepts,” In International Conference on Applications of Natural Language to Information Systems, pp. 407-418, Springer, Cham, 2018.

[25] Mirtalaie, M. A., Hussain, O. K., Chang, E., & Hussain, F. K. (2017). A decision support framework for identifying novel ideas in new product development from cross-domain analysis. Information Systems, 69, 59-80.

[26] Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2017). In search of new product ideas: Identifying ideas in online communities by machine learning and text mining. Creativity and Innovation Management, 26(1), 17-30.

[27] Zhang, M., Fan, B., Zhang, N., Wang, W., & Fan, W. (2021). Mining product innovation ideas from online reviews. Information Processing & Management, 58(1), 102389.

[28] G. May, M. Taisch, and E. Kerga, “Assessment of Sustainable Practices in New Product Development,” Advances in Production Management Systems. Value Networks: Innovation, Technologies, and Management IFIP Advances in Information and Communication Technology, pp. 437–447, 2012.

[29] J. Kim and D. Wilemon, “Focusing the fuzzy front–end in new product development,” R&D Management, vol. 32, no. 4, pp. 269–279, 2002.

[30] J. Kim and D. Wilemon, “Strategic issues in managing innovation’s fuzzy front-end,” European Journal of Innovation Management, vol. 5, no. 1, pp. 27–39, 2002.

[31] S. H. Aschehoug, C. Boks, and S. Støren, “Environmental information from stakeholders supporting product development,” Journal of Cleaner Production, vol. 31, pp. 1-13, 2012.

[32] J. Cagan, C. M. Vogel. Creating breakthrough products: Innovation from product planning to program approval. FT Press, 2002.

[33] M. Crawford, A. Di Benedetto. "New Products Management. 9th.", New York, NY: McGraw-Hill Irwin, 2008.

[34] J. Tidd, J. Bessant, K. Pavitt. Innovation management. Willey, New York, 2001.

[35] V. Krishnan and K. T. Ulrich, “Product Development Decisions: A Review of the Literature,” Management Science, vol. 47, no. 1, pp. 1–21, 2001.

[36] A. Alblas, K. Peters, and J. C. Wortmann, “Fuzzy sustainability incentives in new product development,” International Journal of Operations & Production Management, vol. 34, no. 4, pp. 513–545, 2014.

[37] Wolf, RA (1994). Organizational innovation:Review, critique and suggested research directions. Journal of Management Studies, 31(3), 405–431.

[38] D. Dimancescu, K. Dwenger. “Smoothing the product development path”, Management Review, vol. 85(1), pp. 36, 1996.

[39] A. Griffin, “PDMA research on new product development practices: Updating trends and benchmarking best practices”, Journal of Product Innovation Management, vol. 14, no. 6, pp. 429–458, 1997.

[40] R. Cooper, "Benchmarking new product performance: Results of the best practices study", European Management Journal, vol. 16(1), pp. 1-17, 1998.

[41] Trott, P. (2017). Innovation management and new product development, 6th Edition, Pearson.

[42] A. Riel, M. Neumann, and S. Tichkiewitch, “Structuring the early fuzzy front-end to manage ideation for new product development,” CIRP Annals - Manufacturing Technology, vol.62, no. 1, pp. 107–110, 2013.

[43] Q. Zhang and W. J. Doll, “The fuzzy front end and success of new product development: a causal model,” European Journal of Innovation Management vol. 4, no. 2, pp. 95–112, 2001.

[44] C. T. Su, Y. H. Chen, & D. Y. Sha, “Linking innovative product development with customer knowledge: a data-mining approach,” Technovation, vol. 26, no. 7, pp. 784-795, 2006.

[45] F. Giones, F. and P. Oo, “How crowdsourcing and crowdfunding are redefining innovation management,” in Revolution of Innovation Management, pp. 43–70, 2017.

[46] N. Escoffier, N. Tournois, and B. Mckelvey, “Using crowdsourcing to increase new product’s market value and positive comments for both the crowd involved and customers,” International Journal of Innovation Management, vol. 22, no. 2, pp. 1–28, 2018.

[47] W. D. Hoyer, R. Chandy, M. Dorotic, M. Krafft, and S. S. Singh, “Consumer cocreation in new product development,” Journal of Service Research vol. 13, no. 3, pp. 283–296, 2010.

[48] M. W. G. Rocha, A. S. O. Yu, and P. T. de Souza Nascimento, “Crowdsourcing in the fuzzy front end of innovation,” in Proceedings of 2014 Portland International Conference on Management of Engineering & Technology (PICMET’14), pp. 830–839, 2014.

[49] S. Thomke and E. Von Hippel, “Customers as innovators: a new way to create value,” Harvard Business Review, vol. 80, no. 4, pp. 74–85, 2002.

[50] H. Zhang, and W. Chen, “Crowdfunding technological innovations: Interaction between consumer benefits and rewards”, Technovation, vol. 84, pp.11-20, 2019.

[51] A. Natalicchio, A. M. Petruzzelli, and A. C. Garavelli, “Innovation problems and search for solutions in crowdsourcing platforms–A simulation approach”, Technovation, vol. 64, pp.28-42, 2017.

[52] M. Klein, A.C.B. Garcia, “High-speed idea filtering with the bag of lemons,” Decision Support Systems, vol. 78, pp.39-50, 2015.

[53] D. Dellermann, N. Lipusch, M. Li, “Combining Humans and Machine Learning: A Novel Approach for Evaluating Crowdsourcing Contributions in Idea Contests,” In: Multikonferenz Wirtschaftsinformatik (MKWI). Lüneburg, Germany, 2018.

[54] V. Banken, Q. Ilmer, I Seeber, and S. Haeussler, “A method for Smart Idea Allocation in crowd-based idea selection,” Decision Support Systems, vol. 124, p.113072, 2019.

[55] T. C. Dinh, H. Bae, J. Park, and J. Bae, “A framework to discover potential ideas of new product development from crowdsourcing application,” International Conference on Computer, Networks, Systems, and Industrial Applications, 2012.

[56] J. Ma, Y. Lu, and S. Gupta, “User innovation evaluation: Empirical evidence from an online game community,” Decision Support Systems, vol. 117, pp.113-123, 2019.

[57] Ko, N., Jeong, B., Choi, S., & Yoon, J. (2017). Identifying product opportunities using social media mining: application of topic modeling and chance discovery theory. IEEE Access, 6, 1680-1693.

[58] Jeong, B., Yoon, J., & Lee, J. M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. International Journal of Information Management, 48, 280-290.

[59] X. Luo and S. Du, “Exploring the relationship between corporate social responsibility and firm innovation,” Marketing Letters, vol. 26, no. 4, pp. 703–714, 2014.

[60] D. J. Teece, “Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance,” Strategic Management Journal, vol. 28, no. 13, pp. 1319–1350, 2007.

[61] M. C. Claudy, M. Peterson, and M. Pagell, “The Roles of Sustainability Orientation and Market Knowledge Competence in New Product Development Success,” Journal of Product Innovation Management, vol. 33, pp. 72–85, 2016.

[62] K. S. Swan and M. Luchs, “From the Special Issue Editors: Product Design Research and Practice: Past, Present and Future,” Journal of Product Innovation Management, vol. 28, no. 3, pp. 321–326, 2011.

[63] D. Pujari, Eco-innovation and new product development: understanding the influences on market performance. Technovation, vol. 26, no. 1, pp. 76-85, 2006.

[64] S. Du, G. Yalcinkaya, and L. Bstieler, “Sustainability, Social Media Driven Open Innovation, and New Product Development Performance,” Journal of Product Innovation Management, vol. 33, pp. 55–71, 2016.

[65] R. M. Dangelico and D. Pujari, “Mainstreaming Green Product Innovation: Why and How Companies Integrate Environmental Sustainability,” Journal of Business Ethics, vol. 95, no. 3, pp. 471–486, 2010.

[66] P. H. Driessen, B. Hillebrand, R.A.W. Kok, T.M.M. Verhallen, “Green new product development: the pivotal role of product greenness,” IEEE Trans Eng Manage, vol. 60, no. 2, pp. 315–326, 2013.

[67] C. Moral, A. de Antonio, R. Imbert, and J. Ramírez. "A survey of stemming algorithms in information retrieval." Information Research: An International Electronic Journal, vol. 19(1) 2014.

[68] L. Liu, J. Kang, J. Yu, and Z. Wang, “A comparative study on unsupervised feature selection methods for text clustering,” in 2005 International Conference on Natural Language Processing and Knowledge Engineering, pp. 597–601, IEEE, 2005.

[69] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, “A review of machine learning algorithms for text-documents classification,” Journal of Advances in Information Technology, vol. 1, no. 1, pp. 4–20, 2010.

[70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[71] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[72] X. Wang, J. Wen, S. Alam, Z. Jiang, and Y. Wu, “Semi-supervised learning combining transductive support vector machine with active learning,” Neurocomputing, vol. 173, pp. 1288–1298, 2016.

[73] C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, pp. 273–297, 1995.

[74] K. Crammer and Y. Singer, “On the algorithmic implementation of multiclass kernel-based vector machines, ”Journal of Machine Learning Research, vol. 2, pp. 265–292, 2001.

[75] Y. Lee, Y. Lin, and G. Wahba, “Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data,” Journal of the American Statistical Association, vol. 99, no. 465, pp. 67–81, 2004.

[76] T. Joachims, “Transductive inference for text classification using support vector machines,” in Proceedings of the Sixteenth International Conference on Machine Learning (ICML), pp. 200–209, 1999.

[77] V. Sindhwani and S. S. Keerthi, “Large scale semi-supervised linear SVMs,” in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 477–484, 2006.

[78] S. S. Keerthi and D. DeCoste, “A modified finite Newton method for fast solution of large scale linear SVMs,” Journal of Machine Learning Research, vol. 6, pp. 341–361, 2005.

[79] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, “Foundations on imbalanced classification,” in Learning from Imbalanced Data Sets, pp. 19–46, Springer, 2018.

[80] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[81] S. Maldonado, J. López, and C. Vairetti, “An alternative SMOTE oversampling strategy for high-dimensional datasets,” Applied Soft Computing, vol. 76, pp. 380–389, 2019.

[82] L. Leydesdorff, “Betweenness centrality as an indicator of the interdisciplinary of scientific journals,” J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 9, pp. 1303–1319, 2007.

[83] D. F. Lezzi, “Centrality measures for text clustering,” Commun. Stat. - Theory Methods, vol. 41, no. 16–17, pp. 3179–3197, 2012.