Science of the Total Environment 673 (2019) 402–413

Contents lists available at ScienceDirect

Science of the Total Environment

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

Innovation hotspots in food waste treatment, biogas, and anaerobic digestion technology: A natural language processing approach

Djavan De Clercq [a], Zongguo Wen [a,*], Qingbin Song [b]

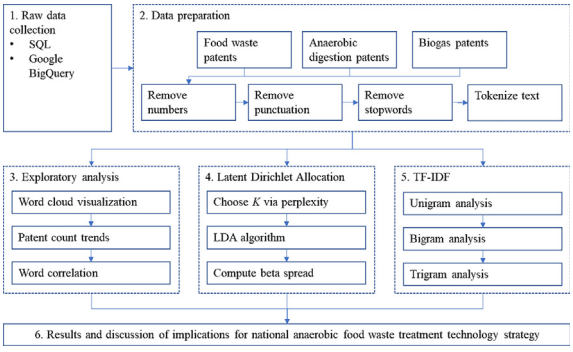
[a] State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China

[b] Macau Environmental Research Institute, Macau University of Science and Technology, Macao

HIGHLIGHTS

- Latent Dirichlet Allocation used to identify topics present in food waste, biogas, and AD patents
- TF-IDF applied to gauging emerging technology concepts across various years of published patents
- Policy implications with regard to technology selection are proposed based on the analysis.
- The entire data and code behind the analysis are open-source and available online.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 February 2019
Received in revised form 2 April 2019
Accepted 3 April 2019
Available online 7 April 2019

Editor: Huu Hao Ngo

Keywords:
Natural language processing
Latent Dirichlet Allocation
TF-IDF
Food waste
Biogas
Anaerobic digestion

ABSTRACT

The objective of this study is to apply natural language processing to identifying innovative technology trends related to food waste treatment, biogas, and anaerobic digestion. The methodology used involved analyzing large volumes of text data mined from 3186 patents related to these three fields. Latent Dirichlet Allocation and the perplexity method were used to identify the main topics which the patent corpora were comprised of and which technological concepts were most associated with each topic. In addition, term frequency-inverse document frequency (TF-IDF) was used to gauge the "emergingness" of certain technical concepts across the patent corpora in various years. The key results were as follows: (1) perplexity computations showed that a 20 topic models were feasible for these patent corpora; (2) topics were identified, providing an accurate picture of the patenting landscape in the analyzed fields; (3) TF-IDF analysis on unigrams, bigrams, and trigrams, supplemented with network graph analysis, revealed emerging technology trends in each year. This study has important implications for governments who need to decide where to invest resources in anaerobic food waste treatment.

1. Introduction

1.1. Anaerobic food waste technology roadmapping difficulties

Governments, agencies, and companies around the world face the problem of "technology roadmapping", which involves setting clear technology targets for an industry based on critical stakeholder

* Corresponding author.
E-mail address: wenzg@tsinghua.edu.cn (Z. Wen).

https://doi.org/10.1016/j.scitotenv.2019.04.051
0048-9697/

requirements of a technology system (Aleina et al., 2017). By providing intelligence on emerging technologies, good technology roadmaps can assist governments and industry to make wise investment decisions and remain competitive in selected technology fields (Lahoti et al., 2018). This research addresses the problem of technology selection in the anaerobic food waste treatment industry.

Unfortunately, it is easy for decision-makers in government and industry to get lost in the plethora of information contained in academic papers, patent texts, and industry reports. Information on appropriate technology for anaerobic food waste treatment can be conflicting, and the methods used to arrive at certain conclusions are sometimes not reproducible, as code/original is not provided in many studies.

For instance, some studies have focused on lab-scale experimentation to determine optimal conditions for anaerobic digestion (AD) of food waste. One example is Deepanraj et al. (2017), which investigated the influence of process parameters (such as TS, pH, temperature, C/N ratio, and ultrasonic pretreatment) on biogas production during the digestion of food waste. Another example is Maragkaki et al. (2018), which conducted lab-scale experiments and found that co-digestion of sewage sludge with food waste, grape residues, crude glycerol, cheese whey and sheep manure increased biogas methane content by 4.5%–8.5%. Similar studies investigating a host of AD-related parameters have been conducted in recent years (Cheng et al., 2018; Kuczman et al., 2018; W. Li et al., 2018; Liu et al., 2016; Menon et al., 2017; Nguyen et al., 2017; Nie et al., 2017; Rajagopal et al., 2017; Ye et al., 2018; Zhang et al., 2017a, 2017b; W. Zhang et al., 2017).

Other studies have focused on life-cycle assessment (LCA) to determine the most environmentally friendly methods to treat food waste. For instance, Thyberg and Tonjes (2017) found that AD food waste treatment offered the fewest environmental burdens, and that incineration performed better than composting. On the other hand, Gao et al. (2017) found that incineration actually had a worse environmental impact than composting. Tong et al. (2018) found that AD followed by composting digestate is best in most LCA impact categories. Laso et al. (2018) found that incineration had the lowest environmental impact out of several surveyed technologies. Cristóbal et al. (2016) applied data envelopment analysis (DEA) and LCA, and found that no specific food waste treatment system was superior across environmental impact categories. On the other hand, Edwards et al. (2017) found that anaerobic co-digestion of sewage sludge and municipal food waste and the least environmental impacts across all categories. Another study by Woon et al. (2016) found that turning food waste into biogas fuel as a petrol substitute was the most advantageous.

Unfortunately, the results from the studies mentioned above can be difficult to generalize to sector-wide technology planning from a policy standpoint. For instance, experimental results do not always generalize to full-fledged industrial AD facilities; in fact, parameters such as waste substrate mixture deemed optimal for biogas production in experiments might actually be “highly improper for industrial applications” (Matuszewska et al., 2016). Moreover, LCA studies often give conflicting results and suffer from high uncertainty due to the differences in system boundaries, scale, and technology types (Brunklaus et al., 2018). Moreover, LCA findings are difficult to interpret since technologies perform differently across various environmental impact categories (Angelo et al., 2017).

1.2. Text analysis of patents can provide valuable information

In departure from experimental and LCA-based approaches, this research provides policymakers with insights into food waste treatment technology innovations from a patent perspective.

Technology patents contain a wealth of information which can assist scientists, engineers, and corporate/political decision makers throughout the inventive process (Madani and Weber, 2016). According to the World Intellectual Property Organization, 90% to 95% of inventions can be found in patent documents (Souili et al., 2015), making patent texts an important resource for understanding the evolution of technologies over time. Patent documents contain (1) patent text data and (2) patent metadata. Text data includes text from the patent's title, abstract, background/summary, detailed description, and claims. Metadata includes fields such as the patent inventor, applicant, date of issue, assignee, patent examiner, and so on.

Patent text data can provide useful information about the technological trends in industries (in this case, food waste, anaerobic digestion and biogas), but parsing through such text consumes significant time and resources. As a result, text analysis methods and natural language processing (NLP) have gained widespread recognition for their ability to garner insights from patent corpora. Applications of NLP include sentiment analysis, topic segmentation/recognition, machine translation, and relationship extraction.

Previous studies have applied NLP to patent data for technology forecasting. For instance, Lee et al. (2018) employed feed-forward multilayer neural networks to assess the value of patents and build an indicator system to evaluate a technology's “emergingness” over time. Kyebambe et al. (2017) proposed an algorithm capable of clustering similar technologies based on patent feature vectors and predicting emerging technologies at least a year before they emerge. Other studies in recent years have also focused on technology forecasting. For instance, Joung and Kim (2017) monitored emerging technologies for technology planning using technical keyword based analysis. Similarly, Song et al. (2018) identified promising technologies based on retrospective feature analysis and prospective needs analysis on outlier patents.

In addition to studies on technology forecasting, other studies have sought to classify patents into technology categories. Zhang (2014) proposed an interactive patent classification algorithm based on multi-classifier fusion and active learning. Wu et al. (2016) developed an automatic patent quality analysis system capable of clustering previously published patents according to their quality, based on self-organizing maps and support vector machine. Venugopalan and Rai (2015) built a classifier based on document-term frequency and topic modeling in order to categorize 10,201 patents about solar photovoltaics by technology area. It was able to separate relevant from irrelevant patents with an accuracy of 98% according to market/product areas.

1.3. NLP techniques have not been applied to food waste, AD, and biogas patents

Natural language processing methods have not yet been applied to extracting useful information from patents at the intersection of food waste treatment, anaerobic digestion, and biogas. Although bibliometric analyses of trends in anaerobic food waste have been conducted (Ren et al., 2018), these have focused largely on qualitative manual reading of literature, rather than text mining approaches.

The study addresses the following research question: what are the trends in patenting activity related to food waste, anaerobic digestion, and biogas? To answer this question, we apply Latent Dirichlet Allocation (LDA) and term frequency-inverse document frequency (TF-IDF) to (1) identifying the topics covered in these patents and (2) identifying emerging technology concepts in the patent corpus. This research is particularly timely as governments around the world seek to incorporate anaerobic food waste treatment technology into national strategies to tackle food waste (De Clercq et al., 2017).

Section 2 introduces the NLP methodology used in this study. Section 3 presents the results and provides discussion of the key findings. Concluding remarks are made in Section 4, and the Supplementary Information provides the entire dataset and the code behind the analysis to ensure full reproducibility of this research.

2. Methodology

The methodology in this study is based on the flowchart in Fig. 1. Firstly, raw data was collected based on SQL query of an open-source

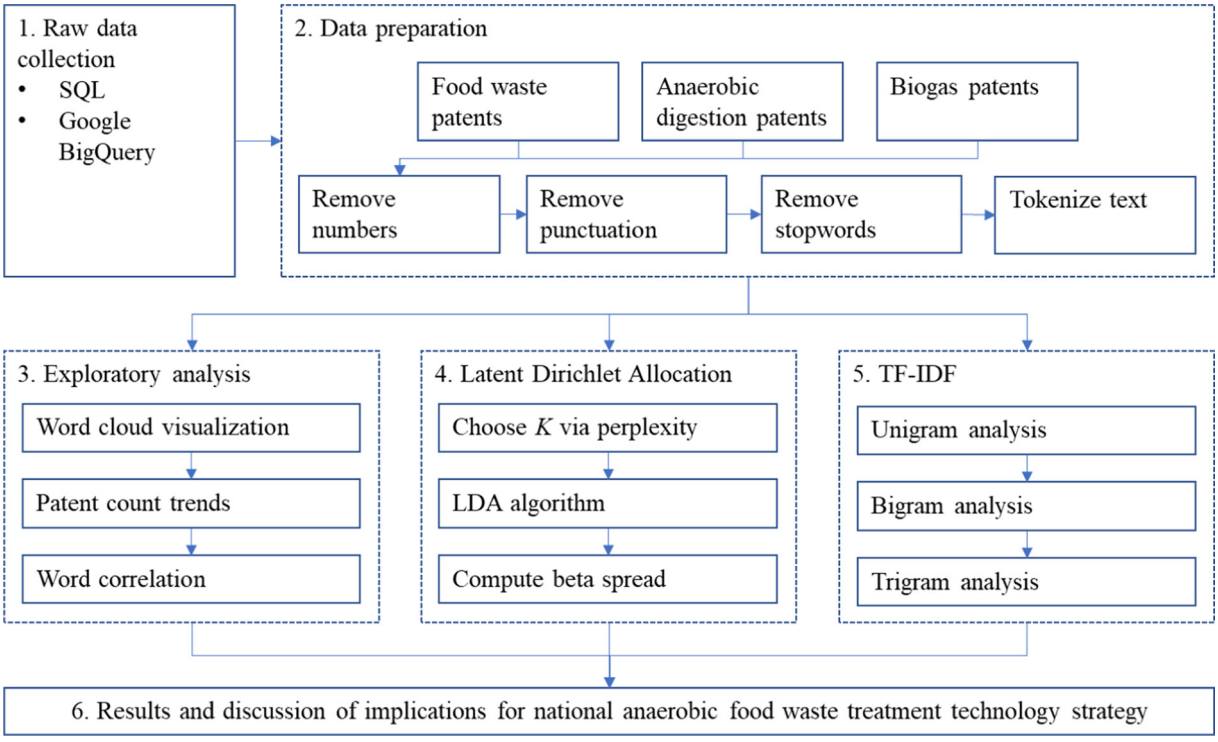


Fig. 1. Overview of methodology.

patent database hosted on Google Cloud (1). Secondly, data was prepared based on several preprocessing steps outlined in (2). Thirdly, patent data was analyzed based on (3), (4), and (5). The results of these findings are the foundation for the discussion presented in (6). The entire dataset and code used in this study is provided in the Supplementary information for full reproducibility.

Each sub-section below corresponds to the numbered box in Fig. 1.

2.1. Raw data collection

Patents related to food waste, anaerobic digestion, and biogas were retrieved via SQL queries of these terms on a patent database hosted on Google Cloud. SQL queries were used for their convenience in formulating queries with regular expressions (regex); this made obtaining large volumes of relevant patent text data easier. Since patents related to food waste, biogas and anaerobic digestion do not have clearly defined technology categories, regex statements with SQL were essential for data acquisition. Other studies, which analyze patents pre-categorized in broad technology spheres, may not require this approach; for instance Chen (2017), which broadly studied "utility patents". SQL queries searching for keywords have been used in studies such as Venugopalan and Rai (2015), which used keyword searches to retrieve patents on solar photovoltaic technology.

The Google BigQuery interface was used to retrieve the data with SQL. The underlying data comes from the United States Patent & Trademark Office's PatentsView database, which contains patent text data and metadata on millions of patents filed through that agency over the last few decades. The database contains patents filed not only by US-based inventors, but from inventors around the world (Germany, Japan, etc.). The retrieved data was downloaded as a .csv file; the entire dataset is provided in the Supplementary information for the reader's convenience.

2.2. Data preparation

Data was prepared before applying NLP algorithms. For instance, numbers and punctuation were removed, as they do not contribute useful information to emerging trends in words and phrases related to technology innovation. In addition, stop words were removed: these are words that are not useful for an analysis, typically very common words such as "the", "of", "to", and so on. In addition, additional stop words were defined, such as "invention", "patent", and "document", since these words occur frequently in patent data. Lastly, the text was tokenized, which entails converting a string of text to a table with one token per row. A "token" is a meaningful unit of text, such as a word or phrase, that we are interested in using for analysis. The entire data preparation process is documented in the GitHub online code supplement: https://github.com/djavandeclercq/FoodWasteBiogasADNLP.

2.3. Exploratory analysis and phi coefficient word correlation

Prior to LDA and TF-IDF analysis, basic exploratory analysis was conducted on the processed text data. Analyses included word counts (Supplementary information, S1), patent number trends over the years (Supplementary information S2), and pairwise word correlation analysis.

Word counts were visualized in a word cloud to verify that the patents retrieved via SQL query were indeed relevant to the topics of interest. Patent numbers over the years were visualized for each field in order to demonstrate trends in patenting activity. Lastly, pairwise word correlation analysis was conducted to examine how often certain words appear together relative to how often they appear separately.

Pairwise word correlations were computed based on the phi coefficient, which is a common measure for binary correlation. This coefficient computes how likely it is that either both word X and Y appear, or neither do, than that one appears without the other (Silge and

Table 1
Values used to compute the phi coefficient.

	Contains word Y	Does not contain word Y	Total
Contains word X	n_{11}	n_{10}	$n_{1\cdot}$
Does not contain word X	n_{01}	n_{01}	n_{11}
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	n

Robinson, 2017). Given Table 1:where n_{11} represents the quantity of documents where word X and word Y co-occur, n_{00} represents the quantity where neither appear, and n_{10} and n_{01} indicate where one word appears without the other. The phi coefficient is computed as:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0} \cdot n_{\cdot 0}n_{\cdot 1}}} \quad (1)$$

and allows us to identify associations between particular words.

2.4. Latent Dirichlet Allocation

In order to extract the topics present in each set of patent documents (food waste, biogas, and anaerobic digestion) this study applied Latent Dirichlet Allocation (LDA), a topic modeling algorithm. LDA is a generative probabilistic model for collections of discrete data such as text corpora, first described in Blei et al. (2003). The main objective of topic modeling is to reveal latent themes across a corpus of documents based on the assumption that the documents are generated by a certain stochastic process (Kim et al., 2019). Bastani et al. (2019) used LDA to conduct topic modeling of consumer complaints in the USA's Consumer Financial Protection Bureau. Another study used LDA to conduct multi-label classification of cardiology records (Pérez et al., 2018). Wang et al. (2018) used LDA for topic analysis of online reviews for two competing products.

In the original model, LDA is used to model a collection of unlabeled documents as a mixture of topics, where each topic is a distribution over fixed terms. LDA has emerged as a popular unsupervised learning model for document and word clustering (Momtazi, 2018).

In the LDA model, we take a $M * V$ co-occurrence table to indicate our patent corpus (e.g. all patents related to food waste), where M is the number of documents and V denotes the size of the vocabulary. This table contains the frequency of occurrences $n(\boldsymbol{w}_i, \boldsymbol{d}_j)$ for word $\boldsymbol{w}_i$ in document $\boldsymbol{d}_j$. LDA assumes that this corpus contains K latent hidden topics ($z_1, z_2, \ldots, z_k$), and that documents in the corpus are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei et al., 2003; H. Li et al., 2018).

Given the parameter α, the probability density can be expressed as (W. Li et al., 2018; H. Li et al., 2018):

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\ldots\theta_K^{\alpha_K-1} \quad (2)$$

where $\Gamma(\cdot)$ is the gamma function and θ is the topic mixture. The joint distribution of θ, topics z, and words w for the given parameters α and β is (Blei et al., 2003):

$$p(\theta, \mathrm{z}, \mathrm{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{p}(\mathrm{z_n}|\theta)\mathrm{p}(\mathrm{w_n}|\mathrm{z_n}, \beta) \quad (3)$$

where N is the number of topics. A document's marginal distribution is obtained by:

$$p(\mathrm{w}|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{z_n}} \mathrm{p}(\mathrm{z_n}|\theta)\mathrm{p}(\mathrm{w_n}|\mathrm{z_n}, \beta)\right)d\theta \quad (4)$$

Lastly, to estimate the topic distribution z for a given document, the posterior distribution is computed as:

$$p(\theta, \mathrm{z}|\mathrm{w}, \alpha, \beta) = \frac{\mathrm{p}(\theta, \mathrm{z}, \mathrm{w}|\alpha, \beta)}{\mathrm{p}(\mathrm{w}|\alpha, \beta),} \quad (5)$$

where $p(\theta,z,w|\alpha,\beta)$ is obtained by Eq. (2) and $p(w|\alpha,\beta)$ is obtained by Eq. (3).

In the context of this study, the objective of LDA is to find the mixture of words that is associated with each of the K topics, and to determine the mixture of topics that describes each document. Patents are then labelled with K topic probabilities, indicating the likelihood that a given patent is related to topic n.

Instead of setting K to an arbitrary value, this study referred to the perplexity measure to determine the appropriate number of topics. Perplexity is a common measure of the probability distribution's predictive ability; appropriate distributions have relatively low perplexity (Wang and Xu, 2018). By convention, language modeling generally uses perplexity as the preferred measure for model evaluation (Hagen, 2018). Due to its high stability, many studies have employed the approach to determine the best LDA parameters (Backenroth et al., 2018; Nabli et al., 2018; Pavlinek and Podgorelec, 2017; Zhang et al., 2016).

Perplexity can be calculated with the following equation (Pavlinek and Podgorelec, 2017):

$$per(D_{test}) = \exp\left(\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right) \quad (6)$$

where D_{test} denotes the held out data, M denotes the number of documents in a collection, w_d denotes the words, and N_d describes the number of words in a given document d.

Lower values of perplexity indicate lower misrepresentation of the words of the test documents by the trained topics. Fig. 2 demonstrates the perplexity of LDA models with varying levels of K; perplexity continues to decrease up until 30 topics, although the scree is reached at about 20 topics. However, perplexity is not the only criteria in deciding on the final value for K; choosing an appropriate K also depends on domain knowledge and human evaluation of whether the words associated with each topic make structural sense. Based on these considerations, this study selected a K value of 20, implying 20 topics; this value had a reasonable value for perplexity, while retaining topic interpretability.

2.5. TF-IDF

TF-IDF is a relevance indicator which evaluates how important a word is in a collection of documents, and his been used in recent studies. Trstenjak et al. (2014) combined TF-IDF with supervised machine learning to conduct text categorization. Wongso et al. (2017) used TF-IDF for news article text classification in the Indonesian language. Erra et al. (2015) conducted a study on the parallel implementation of TF-IDF on graphical processing units (GPUs) to conduct relevance analysis on continuous data streams (including messages, tweets, and sensor-based log files).

Term frequency entails assigning a weight to each term in a document based on the number of occurrences of that term in the document. The weight is then assigned to be equal to the number of occurrences of the term t in the document d. This weighting scheme is denoted $\mathrm{tf}_{t,d}$ with subscripts t and d indicating the term and the document (Trstenjak et al., 2014).

Tf is then multiplied by inverse document frequency in order to attenuate the effect of terms that occur too often in the collection to be useful for differentiating documents. Document frequency df_t is the number of documents in the collection that contain a term t. The document frequency of a term is used to scale its weight by taking the total number of documents N in a collection and defining the *inverse document frequency* of term t by (Manning et al., 2008):

$$idf_t = \log\left(\frac{N}{\mathrm{df_t}}\right) \quad (7)$$

Combining the definitions of term frequency and inverse document frequency results in a composite weight for each term in each document. The tf-idf weighting ascribes to term t a weight in document d

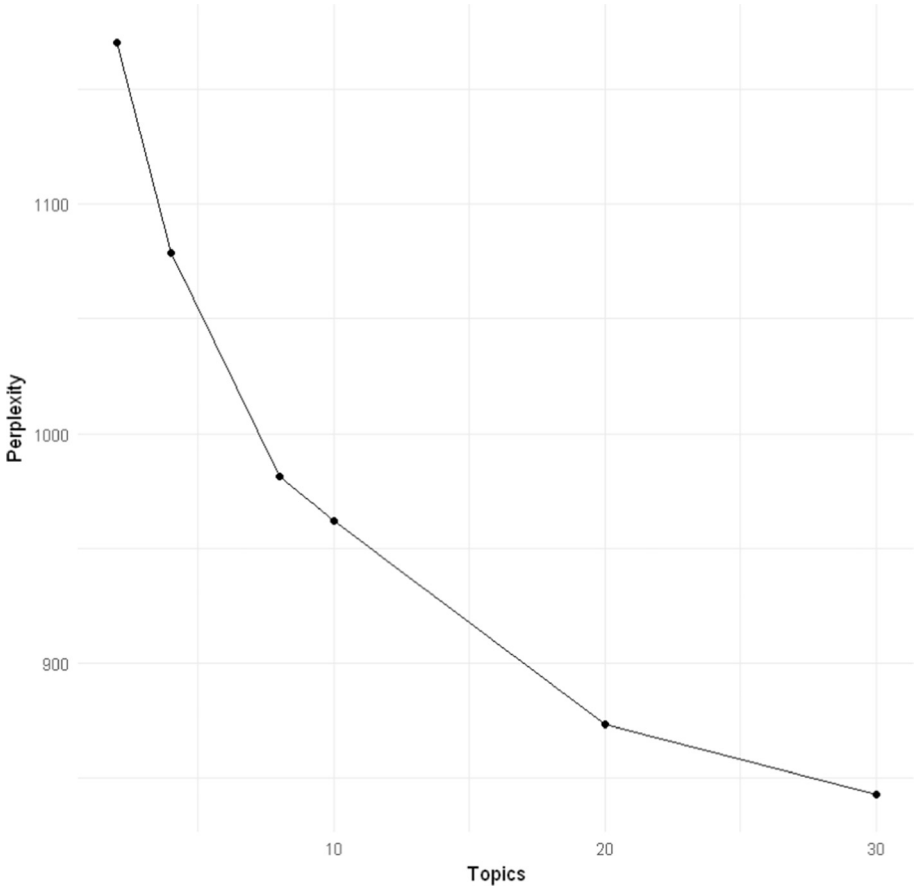


Fig. 2. Perplexity on the validation set depending on the number of topics.

given by

$$\mathrm{tf}_{\mathrm{idf}_{t,d}} = \mathrm{tf}_{t,d} * \mathrm{idf}_{\mathrm{t}} \quad (8)$$

The tf-idf score is: (1) high when *t* occurs frequently within a small number of documents (giving high discriminating power to these documents); (2) lower when *t* occurs infrequently in a document, or occurs in many documents; and (3) lowest when *t* occurs in all documents. In other words, high scores indicate terms that are unique relative to other words in the corpus.

Simply put, the TF-IDF statistic is a measure of how important a word is to a document (patent) in a collection (patent corpus) of documents. The metric allows us to quantify how important various words are in a document that is part of a collection (Silge and Robinson, 2017). In the context of this study, patents in each corpus (food waste, biogas, anaerobic digestion) were separated by year. The objective is to determine the TF-IDF scores for words in patents published during a specific year. This allows us to gauge which technology trends were "emerging" in which particular year for each corpus. The TF-IDF scores were computed for unigrams (one word), bigrams (two-word pieces of text), and trigrams (three-word pieces of text).

2.6. Computation

All the computation in this study was conducted in the Microsoft R Open (MRO) distribution of the R statistical programming language. Jupyter Notebook was used as an IDE, and the entire notebook code behind this study is open-source and can be found online: https://github.com/djavandeclercq/FoodWasteBiogasADNLP.

3. Results & discussion

3.1. Exploratory analysis

S1 (Supplementary information) visualizes the words that occurred the most frequently across patents related to food waste, biogas, and anaerobic digestion. These word clouds verify that the patents retrieved via SQL query of the database were indeed relevant to the analysis. Typical words in the food waste patent corpus included: "food", "waste", "organic", "disposal", "polymer", "fermentation", and "temperature". Typical words in the biogas patent corpus included "anaerobic", "gas", "hydrogen", "sludge", "digester", and "methane". Typical words in the anaerobic digestion patent corpus included "digestion", "sludge", "water", "methane", "biomass", "vessel", and "organic". This confirmed the overall relevance of the retrieved text data. S2 visualizes patenting activity in each category since 1990. For all three categories, patenting was relatively stable until 2008, after which patenting activity increased rapidly.

3.2. Phi-coefficient word correlation

The correlation networks visualized in Fig. 3 demonstrate which keywords occur more often together than other keywords. Notice the presence of both larger, interconnected clusters and smaller, distinct clusters.

In the food waste patent corpus, several distinct clusters emerged. For example, one cluster has the words "citrus", "limonene", and "pectin"; these words can be found in patents related to the conversion of citrus waste into valuable products. Pourbafrani et al. (2010) have also investigated the production of limonene and pectin from citrus waste

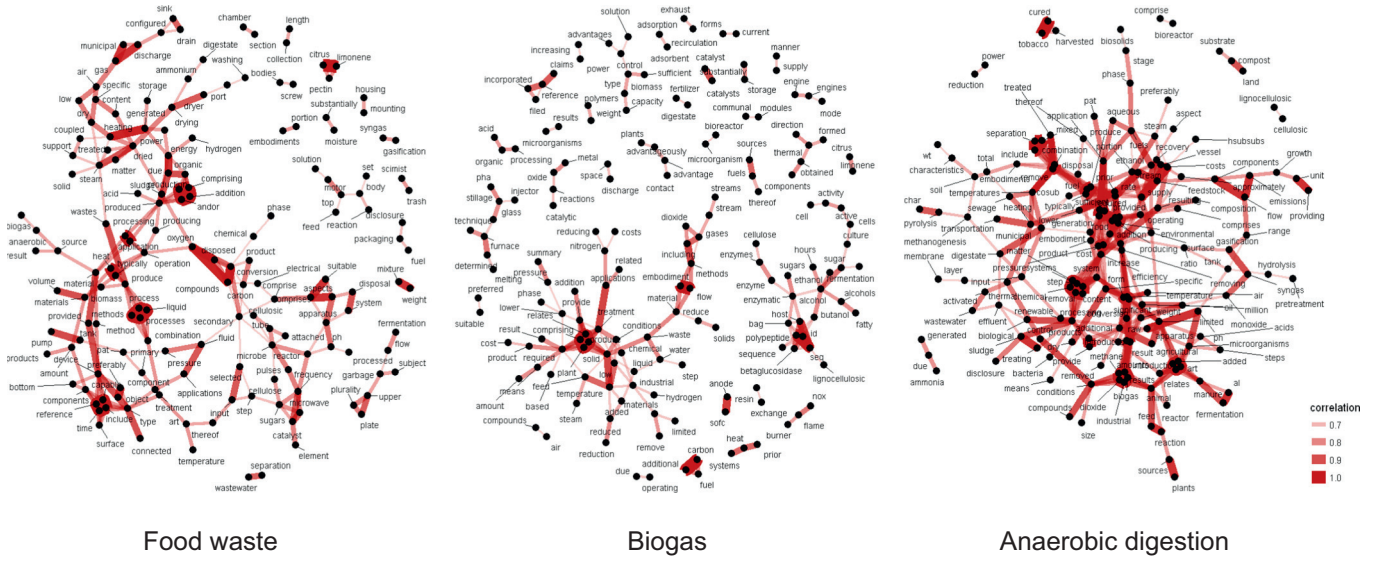


Fig. 3. Pairs of words in the patent corpora that show at least a 0.60 correlation of appearing within the same 4-line section.

via an integrated process, finding that one ton of citrus waste resulted in 39.64 l of ethanol, 45 m^3 of biogas, 8.9 l of limonene, and 38.8 kg of pectin. Similarly, Lotito et al. (2018) tested mesophilic anaerobic digestion for the treatment of citrus waste, showing effective methane production and no limonene inhibition. Negro et al. (2016) found that anaerobic digestion and was suitable for treating citrus waste following limonene extraction. Other distinct clusters included: (1) "separation" and "wastewater"; (2) "chamber" and "section"; (3) "syngas" and "gasification".

In the biogas patent corpus, the correlation network revealed even more distinct clusters of words, possibly due to the more limited technical scope of biogas compared to food waste. As in the food waste corpus, the words "citrus" and "limonene" formed a distinct cluster. Other notable clusters included: (1) "metal", "oxide", "reactions", "catalytic"; (2) "bioreactor", "microorganism"; (3) "adsorption", "recirculation", (4) "burner", "flame", "NOx"; (5) "cellulose", "enzymes".

In the anaerobic digestion patent corpus, the network showed less distinct clusters than for biogas-related patents. Some of the clusters included: (1) "power", "reduction"; (2) "cured", "tobacco", "harvested"; (3) "substrate", "compost", "land"; (3) "char", "pyrolysis"; (4) "separation", "mixed". Recent studies validating the relevance of these clusters include Liu et al. (2015), which investigated the co-digestion of tobacco waste with different agricultural biomass feedstocks, finding that the methane yield of tobacco stalk was 0.163 m^3 CH_4 • kg VS^{-1}. Similarly, González-González et al. (2013) demonstrated the effectiveness of anaerobic digestion applied to treating tobacco plant.

In addition to the networks visualized in Fig. 3, the R code accompanying this study allows the users to select particular words of interest and find other words which are most associated with them. For instance, Fig. 4 demonstrates this functionality for three arbitrarily chosen words of interest: "cellulose", "citrus", and "limonene". The figure shows, for instance, that "cellulose" was highly correlated with "enzymes", "sugars", and "biomass". Such word correlations can direct users of this tool to novel research directions that may provide insight into the best approach to dealing with certain types of biowaste. For instance, Wyman et al. (2018) investigated a lignocellulosic waste valorization strategy through enzyme and biogas production, finding that the *Pleurotus eryngii* fungus engendered higher biogas and enzyme production. These word correlations were also useful in discovering research by Malayil and Chanakya (2016), which investigated fungal enzyme cocktail treatment of biomass for higher biogas production from leaf litter, concluding that biogas yield increased by 29% using this method.

3.3. Topics identified based on Latent Dirichlet Allocation

The results of the LDA and perplexity computations produced the following results. Firstly, the perplexity analysis showed that 20 topics were appropriate to describe the three sets of patent documents. Secondly, for each topic, the probability (beta) of a certain word belonging to that topic was computed. Fig. 5 demonstrates the top 10 words associated with each topic for the biogas patent corpus in particular. The entire LDA results can be found in the Supplementary information S8 and the accompanying online code. The topic modeling process has identified groupings of words that one can understand as human readers of these description fields.

For instance, Fig. 5 (biogas patents only) shows that words in topic 7 with a high probability (beta) of belonging to that topic include "membrane", "membranes", "gas", and "separation", indicating that this topic is likely related to gas purification via membrane technology. In addition, the top words in topic 12 include "gas", "temperature", "pressure", "flow", and "measuring", suggesting that this topic is closely related to measurement of various biogas properties. Words in topic 18 with high betas include "water", "media", "treatment", "wastewater", "activated", and "siloxanes", suggesting association with pollutant removal (such as siloxanes) from sludge or wastewater (Dewil et al., 2007; Oshita et al., 2014).

One feature of LDA topic modeling is that certain words, such as "gas" and "methane" may be common in several topics. Compared to hard clustering methods, this is an advantage of LDA, in that it allows topics to have some overlap (Silge and Robinson, 2017). Nevertheless, one may wish for topics to be further constrained to a set of especially relevant words, for example by considering the words that had the greatest difference in probability between topic 1 and topic 2. This can be computed based on the log ratio between the two:

$$\log_2\left(\frac{\beta_2}{\beta_1}\right) \tag{9}$$

where β_2 is the probability of a given word occurring in topic 2 and β_1 is the probability of a given word occurring in topic 1. Calculating this log ratio after filtering common words (i.e. only computing this ratio for words with a β value greater than 1/1000 for a given topic can yield words with even more discriminating power).

Fig. 6 provides a good example of this, and shows that discriminatory words for topic 1 include "menstruum", "septum", "methanogens", "upflow", "stirred", and "reactor", indicating that topic 1 is closely

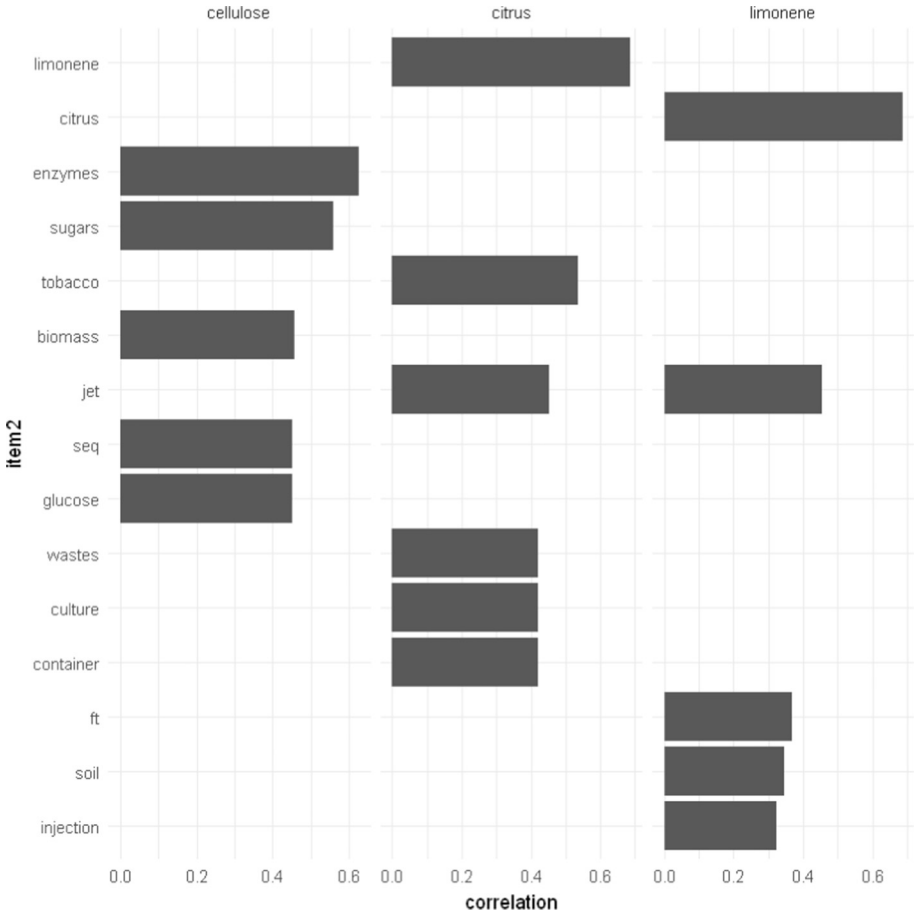


Fig. 4. The top five words from the biogas patent corpus that were most correlated with "cellulose", "citrus", and "limonene" in a 4-line section.

related to specifics of gas production in a digester. Topic 2 words include "regenerative", "dew", "ac" (current), "phototherapy", "inverter", suggesting association with electricity production or transmission. Additional results across the three sets of patent documents are provided in the online code supplement, and users can compute the log ratio for words across various topics of interest. These topics can also be helpful in directed efforts at monitoring trends in innovation. The words in topic 1, for instance, are closely related to research conducted by Reddy et al. (2016), which investigated the performance and emission characteristics of a biogas-fuelled electric generator integrated with a solar concentrated photovoltaic system.

3.4. Emerging technologies identified based on TF-IDF

As described in Section 2, the TF-IDF statistic is a measure of how important a word is to a document (patent) in a collection (patent corpus) of documents. The metric allows us to quantify how important various words are in a document that is part of a collection. Fig. 7 shows the TF-IDF scores for bigrams and trigrams from 2012 to 2017 for the biogas patent corpus specifically. Additional results for the food waste and anaerobic digestion corpora are provided in Supplementary information S3, S4, S5, and in the accompanying online code.

The left panel of Fig. 7 shows bigrams (two words) with the highest TF-IDF in the biogas corpus across several years. For example, technologies characteristic of patenting in 2013 were closely related to concepts such as "rotating container", "methanogenesis tank", "nitrous acid", and "hydrocarbon decomposition".

The right panel of Fig. 7 offers additional granularity by showing trigrams (three words) with the highest TF-IDF in the biogas corpus across several years. For instance, in 2017, they reveal patenting related to concepts including "oxygen transport membrane", "pressure synthesis gas", "micro turbine assembly", and "carbon dioxide absorbing". Other identified trigrams such as "directly injected fuel" are more difficult to attribute to specific concepts without domain knowledge, suggesting that computing TF-IDF scores of higher-order n-grams (i.e. four words or five words) may be informative.

Nevertheless, recent scientific studies underscore the innovation taking place in these technology areas. For instance, regarding "micro turbine assembly", (MosayebNezhad et al., 2019) conducted a technology review and thermodynamic performance study of a biogas-fed micro humid air turbine, showing that the proposed system could achieve an electrical efficiency of 46.6% and a CHP efficiency of 81.2%. As for "carbon dioxide absorbing", recent studies demonstrating activity in this technology sphere include Ferella et al. (2017), which used zeolites as sorbent material for CO_2 capture in biogas upgrading, and Mamun et al. (2016), which researched the use of solid-CaO, CaO-solution, and activated-carbon to remove CO_2 from raw biogas.

In addition to visualizing the top few n-grams as shown in Fig. 7, bigrams and trigrams were also visualized in a network of connected nodes, which provides additional detail about the overall structure of the patent text across all years. In Fig. 8, the nodes represent individual words. The transparency of the edges represents the relative rarity of the bigram (two connected nodes). The arrow directionality demonstrates the direction of the bigram, i.e. which word follows another.

The bigram network in Fig. 8 reveals useful properties of this particular corpus. For example, words such as "sewage", "anaerobic", "biomethane", and "gas" form common centers of nodes, which are followed by other specific concepts. In addition, some words in particular bigrams appear to more isolated, such as "reverse osmosis", "oxygen demand", and "lubricity improver". This network chart allows for

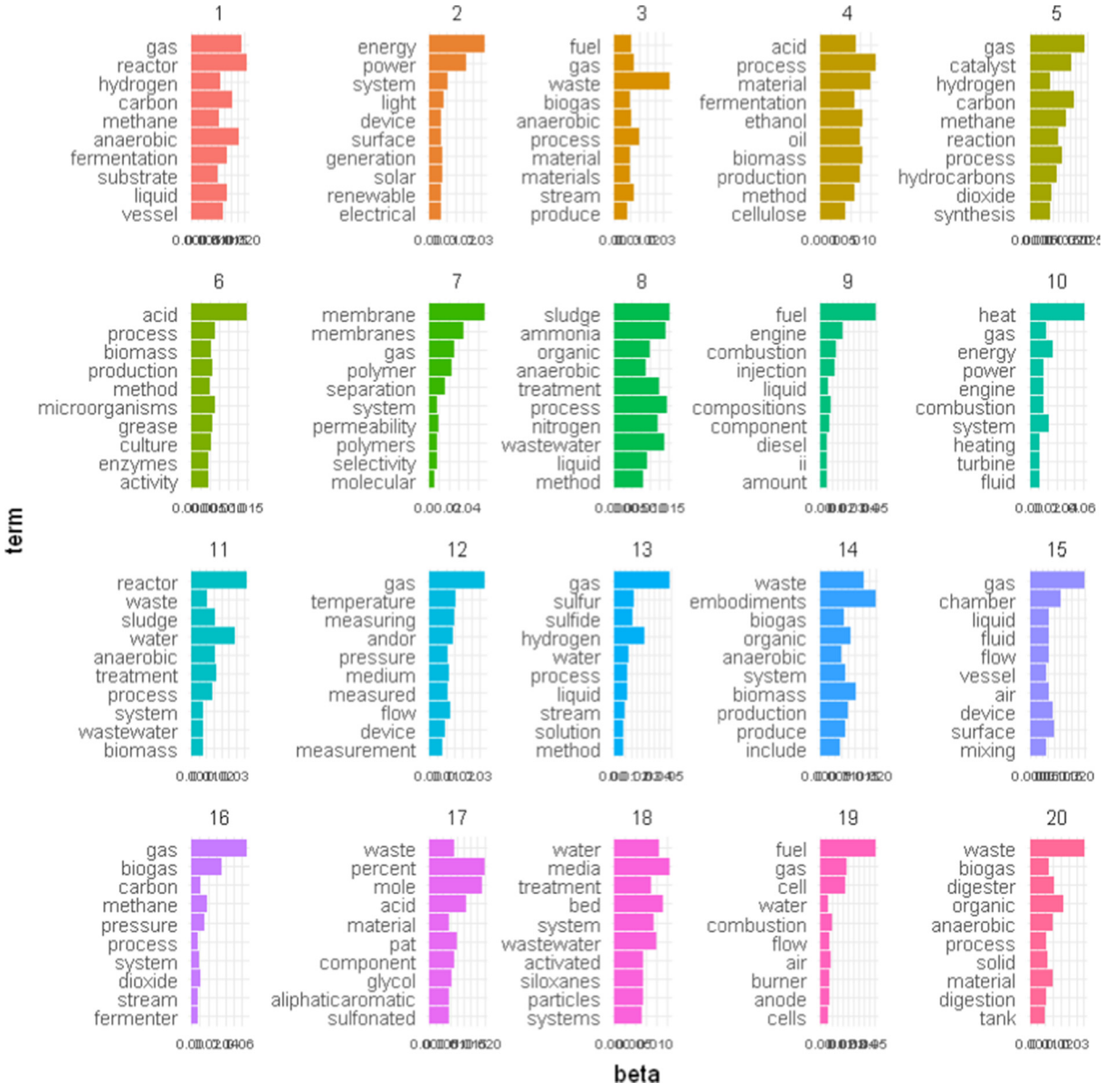


Fig. 5. Top 10 words in each LDA topics for the biogas patent corpus.

isolation of technology concepts that representative of this particular body of patent text.

Several recent studies are closely related to clusters inferred from the bigram network in Fig. 8. Regarding "reverse osmosis", for instance, Zhou et al. (2019) investigated the valorization of biogas slurry with a pilot dual-stage reverse osmosis membrane process, finding that the COD and ammonia rejection rates were 99.41% and 99.11% respectively. Regarding the link the words "aromatic" and "polyimide", recent research (Abd. Hamid et al., 2019) has investigated a miscible-blend polysulfone/polyimide membrane for hydrogen purification from palm oil mill effluent anaerobic fermentation, showing promising results for H_2/CO_2 separation. Regarding "carbonic" and "anhydrase", Fosbøl et al. (2017) conducted design and simulation of rate-based CO_2 capture processes using carbonic anhydrase (CA) applied to biogas, which had implications for biogas upgrading. In their comprehensive review of trends in biogas upgrading technologies and future perspectives, Sahota et al. (2018) also showed that the CA is an emerging approach for dragging CO_2 present in gas into the aqueous phase, where it can be selected by an absorbent.

Both bigram and trigram network graphs were generated for all three sets of patent corpora (food waste, biogas, and anaerobic digestion). They can be seen in Supplementary information S6 and S7 and in the accompanying online code, and readers are encouraged to explore the recent scientific advances associated with the spheres of emerging technologies identified in the analysis.

4. Conclusion

This study addressed the problem of identifying technology trends pertaining to food waste treatment, biogas, and anaerobic digestion, based on text mining methods applied to 3186 patent documents. Phi-coefficient word correlation, Latent Dirichlet Allocation and TF-IDF were used to characterize these trends. The study provided extended upon past literature, which has primarily focused on lab-scale and LCA studies, by offering insights into technology trends based on large text data.

Notable results of the LDA analysis include the following. First, the perplexity measure indicated that 20-topic models were appropriate for the three sets of patent documents. In addition, LDA topics showed

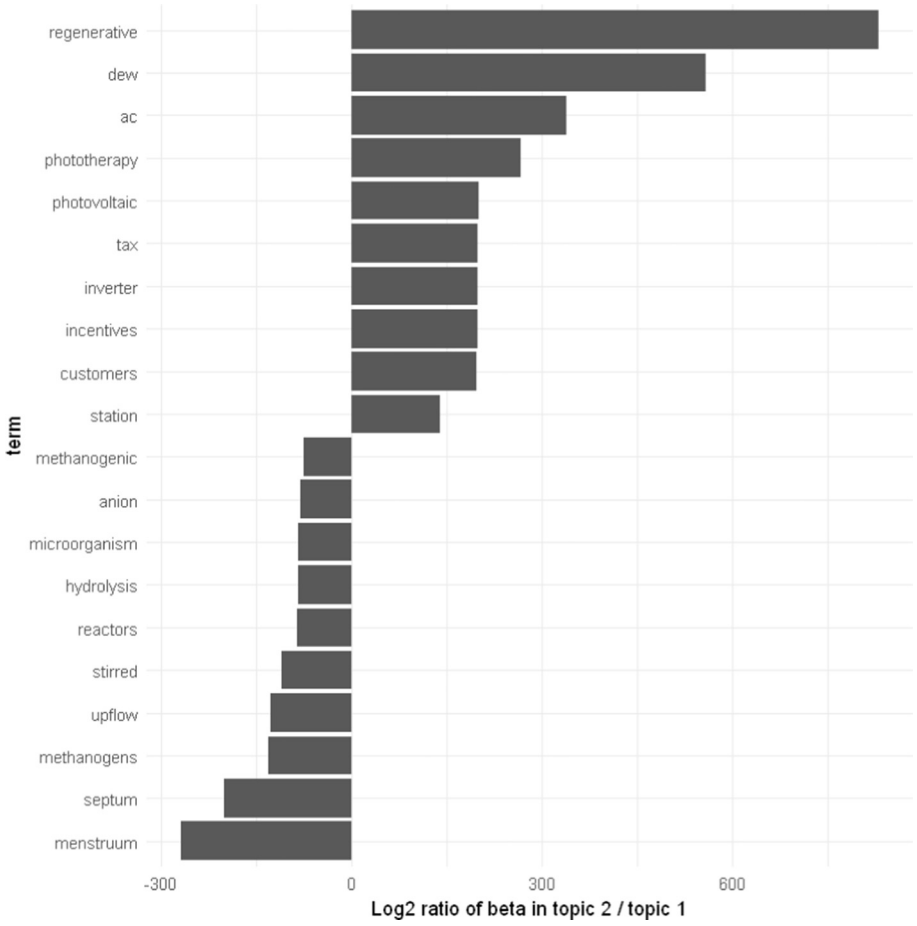


Fig. 6. Words with the greatest difference in β between topic 1 and topic 2 (biogas patent corpus).

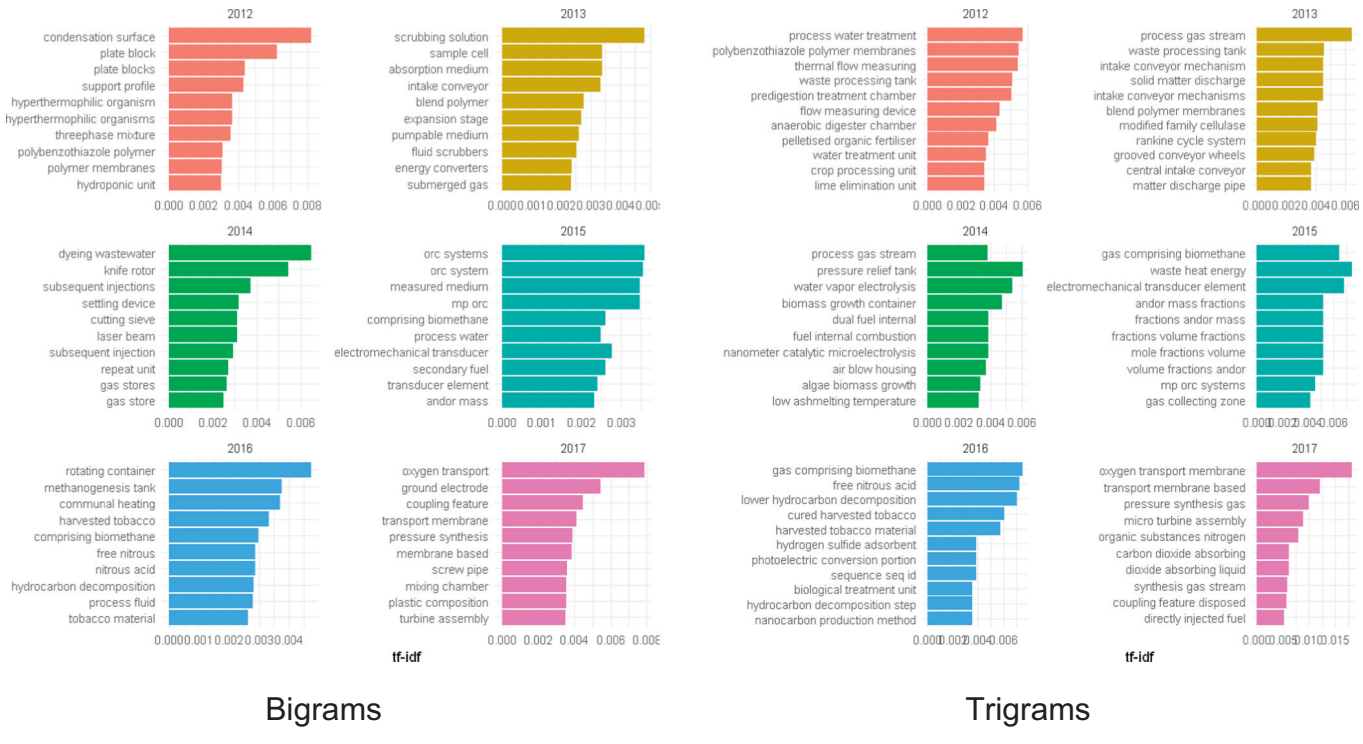


Fig. 7. Highest TF-IDF bigrams and trigrams in each year from 2012 to 2017 for the biogas patent corpus.

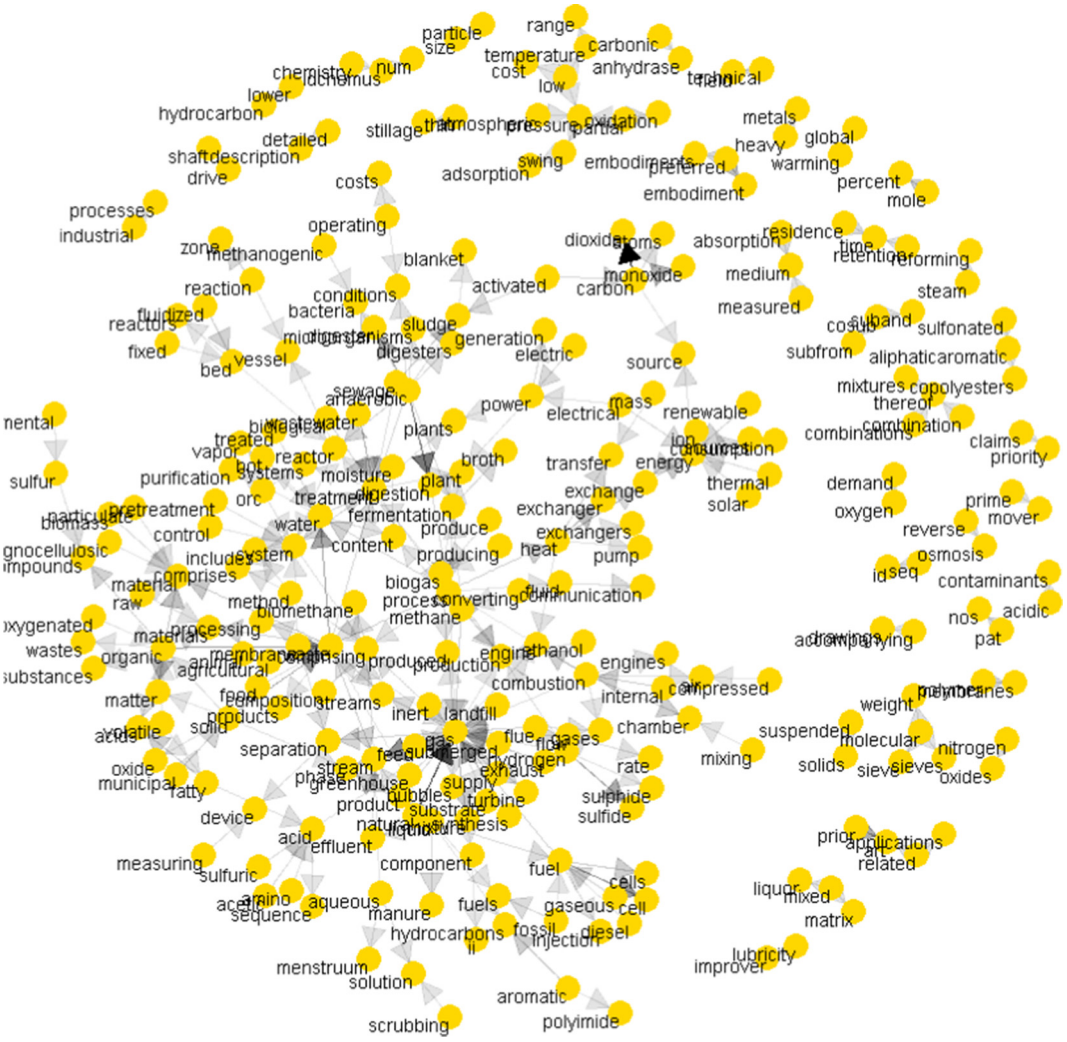


Fig. 8. Common bigrams the biogas patent corpus that occurred at least 100 times.

powerful discriminatory power. For example, one topic in the biogas corpus was clearly associated with pollutant removal (such as siloxanes) from sludge or wastewater, since high-probability words in that topic included "water", "media", "treatment", "wastewater", "activated", and "siloxanes".

Notable results of the TF-IDF results include the following. Characteristic technology concepts for the **food waste** patent corpus in 2016 and 2017 included "inhibitory secondary products", "hydrothermal low temperature", "Fenton reaction catalyst", "biomass pyrolyzing zone", and "concentrated organic waste", among others. Characteristic technology concepts for the **biogas** patent corpus in 2016 and 2017 included: "free nitrous acid", "hydrogen sulfide adsorbent", "biological treatment unit", "nanocarbon production method", "oxygen transport membrane", "pressure synthesis gas", and "micro turbine assembly". Characteristic technology concepts for the **anaerobic digestion** patent corpus in 2016 and 2017 included: "separation composite membrane", "waste processing tank", "polar biomass solution", "waste heat energy", "gas separation composite", and "gas separating layer".

Stakeholders who are deciding to invest significant resources in developing anaerobic food waste treatment technology can benefit greatly from text-mining approaches to relevant patent documents. Coupled with domain expertise and analysis of recent advances in scientific literature, natural language processing-based approaches can provide vital information to technology policy, especially given the fact that patented technologies are often patented with commercialization in mind. Since laboratory-scale experimental studies related to food waste treatment, biogas, and anaerobic digestion technologies may suffer from low scalability towards industrial-scale processes, patent analysis provides vital additional information to policymakers in various regions on how to invest resources wisely into anaerobic food waste treatment. It is suggested that the results from approaches based on text mining be cross-validated with the results from studies adopting other approaches, such as lab-scale experimentation, so that the technology roadmapping process incorporates a broader spectrum of information.

Acknowledgements

The authors gratefully acknowledge the financial support from the "Thirteenth Five-Year" National Key R&D Program of China (2018YFC1903000), National Science Fund for Distinguished Young Scholars of China (71825006) and General Programs (71774099). The responsibility for any error rests solely with the authors. The contents of this paper reflect the views of the authors and do not necessarily indicate acceptance by the sponsors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2019.04.051.

References

Aleina, S.C., Viola, N., Fusaro, R., Saccoccia, G., 2017. Approach to technology prioritization in support of moon initiatives in the framework of ESA exploration technology roadmaps. Acta Astronaut 139, 42–53. https://doi.org/10.1016/j.actaastro.2017.06.029.

Angelo, A.C.M., Saraiva, A.B., Clímaco, J.C.N., Infante, C.E., Valle, R., 2017. Life cycle assessment and multi-criteria decision analysis: selection of a strategy for domestic food waste management in Rio de Janeiro. J. Clean. Prod. 143, 744–756. https://doi.org/10.1016/j.jclepro.2016.12.049.

Backenroth, D., He, Z., Kiryluk, K., Boeva, V., Pethukova, L., Khurana, E., Christiano, A., Buxbaum, J.D., Ionita-Laza, I., 2018. FUN-LDA: a latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. Am. J. Hum. Genet. 102, 920–942. https://doi.org/10.1016/j.ajhg.2018.03.026.

Bastani, K., Namavari, H., Shaffer, J., 2019. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Syst. Appl. 127, 256–271. https://doi.org/10.1016/j.eswa.2019.03.001.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022.

Brunklaus, B., Rex, E., Carlsson, E., Berlin, J., 2018. The future of Swedish food waste: an environmental assessment of existing and prospective valorization techniques. J. Clean. Prod. 202, 1–10. https://doi.org/10.1016/j.jclepro.2018.07.240.

Chen, L., 2017. Do patent citations indicate knowledge linkage? The evidence from text similarities between patents and their citations. J. Inf. Secur. 11, 63–79. https://doi.org/10.1016/j.joi.2016.04.018.

Cheng, H., Hiro, Y., Hojo, T., Li, Y.-Y., 2018. Upgrading methane fermentation of food waste by using a hollow fiber type anaerobic membrane bioreactor. Bioresour. Technol. 267, 386–394. https://doi.org/10.1016/j.biortech.2018.07.045.

Cristóbal, J., Limleamthong, P., Manfredi, S., Guillén-Gosálbez, G., 2016. Methodology for combined use of data envelopment analysis and life cycle assessment applied to food waste management. J. Clean. Prod. 135, 158–168. https://doi.org/10.1016/j.jclepro.2016.06.085.

De Clercq, D., Wen, Z., Gottfried, O., Schmidt, F., Fei, F., 2017. A review of global strategies promoting the conversion of food waste to bioenergy via anaerobic digestion. Renew. Sust. Energ. Rev. 79, 204–221. https://doi.org/10.1016/j.rser.2017.05.047.

Deepanraj, B., Sivasubramanian, V., Jayaraj, S., 2017. Multi-response optimization of process parameters in biogas production from food waste using Taguchi – Grey relational analysis. Energy Convers. Manag. 141, 429–438. https://doi.org/10.1016/j.enconman.2016.12.013.

Dewil, R., Appels, L., Baeyens, J., Buczynska, A., Van Vaeck, L., 2007. The analysis of volatile siloxanes in waste activated sludge. Talanta 74, 14–19. https://doi.org/10.1016/j.talanta.2007.05.041.

Edwards, J., Othman, M., Crossin, E., Burn, S., 2017. Anaerobic co-digestion of municipal food waste and sewage sludge: a comparative life cycle assessment in the context of a waste service provision. Bioresour. Technol. 223, 237–249. https://doi.org/10.1016/j.biortech.2016.10.044.

Erra, U., Senatore, S., Minnella, F., Caggianese, G., 2015. Approximate TF–IDF based on topic extraction from massive message stream using the GPU. Inf. Sci. (Ny). 292, 143–161. https://doi.org/10.1016/j.ins.2014.08.062.

Ferella, F., Puca, A., Taglieri, G., Rossi, L., Gallucci, K., 2017. Separation of carbon dioxide for biogas upgrading to biomethane. J. Clean. Prod. 164, 1205–1218. https://doi.org/10.1016/j.jclepro.2017.07.037.

Fosbøl, P.L., Gaspar, J., Jacobsen, B., Glibstrup, J., Gladis, A., Diaz, K.M., Thomsen, K., Woodley, J.M., von Solms, N., 2017. Design and simulation of rate-based CO2 capture processes using carbonic anhydrase (CA) applied to biogas. Energy Procedia 114, 1434–1443. https://doi.org/10.1016/j.egypro.2017.03.1268.

Gao, A., Tian, Z., Wang, Z., Wennersten, R., Sun, Q., 2017. Comparison between the technologies for food waste treatment. Energy Procedia 105, 3915–3921. https://doi.org/10.1016/j.egypro.2017.03.811.

González-González, A., Cuadros, F., Ruiz-Celma, A., López-Rodríguez, F., 2013. Potential application of anaerobic digestion to tobacco plant. Fuel 113, 415–419. https://doi.org/10.1016/j.fuel.2013.06.006.

Hagen, L., 2018. Content analysis of e-petitions with topic modeling: how to train and evaluate LDA models? Inf. Process. Manag. 54, 1292–1307. https://doi.org/10.1016/j.ipm.2018.05.006.

Hamid, M.A.Abd., Chung, Y.T., Rohani, R., Junaidi, M.U.Mohd., 2019. Miscible-blend polysulfone/polyimide membrane for hydrogen purification from palm oil mill effluent fermentation. Sep. Purif. Technol. 209, 598–607. https://doi.org/10.1016/j.seppur.2018.07.067.

Joung, J., Kim, K., 2017. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technol. Forecast. Soc. Change 114, 281–292. https://doi.org/10.1016/j.techfore.2016.08.020.

Kim, D., Seo, D., Cho, S., Kang, P., 2019. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Inf. Sci. (Ny). 477, 15–29. https://doi.org/10.1016/j.ins.2018.10.006.

Kuczman, O., Gueri, M.V.D., De Souza, S.N.M., Schirmer, W.N., Alves, H.J., Secco, D., Buratto, W.G., Ribeiro, C.B., Hernandes, F.B., 2018. Food waste anaerobic digestion of a popular restaurant in Southern Brazil. J. Clean. Prod. 196, 382–389. https://doi.org/10.1016/j.jclepro.2018.05.282.

Kyebambe, M.N., Cheng, G., Huang, Y., He, C., Zhang, Z., 2017. Forecasting emerging technologies: a supervised learning approach through patent analysis. Technol. Forecast. Soc. Change 125, 236–244. https://doi.org/10.1016/j.techfore.2017.08.002.

Lahoti, G., Porter, A.L., Zhang, C., Youtie, J., Wang, B., 2018. Tech mining to validate and refine a technology roadmap. World Pat. Inf. 55, 1–18. https://doi.org/10.1016/j.wpi.2018.07.003.

Laso, J., Margallo, M., García-Herrero, I., Fullana, P., Bala, A., Gazulla, C., Polettini, A., Kahhat, R., Vázquez-Rowe, I., Irabien, A., Aldaco, R., 2018. Combined application of life cycle assessment and linear programming to evaluate food waste-to-food strategies: seeking for answers in the nexus approach. Waste Manag. 80, 186–197. https://doi.org/10.1016/j.wasman.2018.09.009.

Lee, C., Kwon, O., Kim, M., Kwon, D., 2018. Early identification of emerging technologies: a machine learning approach using multiple patent indicators. Technol. Forecast. Soc. Change 127, 291–303. https://doi.org/10.1016/j.techfore.2017.10.002.

Li, H., Yang, X., Jian, L., Liu, K., Yuan, Y., Wu, W., 2018. A sparse representation-based image resolution improvement method by processing multiple dictionary pairs with latent Dirichlet allocation model for street view images. Sustain. Cities Soc. 38, 55–69. https://doi.org/10.1016/j.scs.2017.12.020.

Li, W., Loh, K.-C., Zhang, J., Tong, Y.W., Dai, Y., 2018. Two-stage anaerobic digestion of food waste and horticultural waste in high-solid system. Appl. Energy 209, 400–408. https://doi.org/10.1016/j.apenergy.2017.05.042.

Liu, C., Li, H., Zhang, Y., Liu, C., 2016. Improve biogas production from low-organic-content sludge through high-solids anaerobic co-digestion with food waste. Bioresour. Technol. 219, 252–260. https://doi.org/10.1016/j.biortech.2016.07.130.

Liu, Y., Dong, J., Liu, G., Yang, H., Liu, W., Wang, L., Kong, C., Zheng, D., Yang, J., Deng, L., Wang, S., 2015. Co-digestion of tobacco waste with different agricultural biomass feedstocks and the inhibition of tobacco viruses by anaerobic digestion. Bioresour. Technol. 189, 210–216. https://doi.org/10.1016/j.biortech.2015.04.003.

Lotito, A.M., De Sanctis, M., Pastore, C., Di Iaconi, C., 2018. Biomethanization of citrus waste: effect of waste characteristics and of storage on treatability and evaluation of limonene degradation. J. Environ. Manag. 215, 366–376. https://doi.org/10.1016/j.jenvman.2018.03.057.

Madani, F., Weber, C., 2016. The evolution of patent mining: applying bibliometrics analysis and keyword network analysis. World Pat. Inf. 46, 32–48. https://doi.org/10.1016/j.wpi.2016.05.008.

Malayil, S., Chanakya, H.N., 2016. Fungal enzyme cocktail treatment of biomass for higher biogas production from leaf litter. Procedia Environ. Sci. 35, 826–832. https://doi.org/10.1016/j.proenv.2016.07.099.

Mamun, M.R. Al, Karim, M.R., Rahman, M.M., Asiri, A.M., Torii, S., 2016. Methane enrichment of biogas by carbon dioxide fixation with calcium hydroxide and activated carbon. J. Taiwan Inst. Chem. Eng. 58, 476–481. https://doi.org/10.1016/j.jtice.2015.06.029.

Manning, C.D., Schütze, H., Raghavan, P., 2008. Preface. Introduction to Information Retrieval. Cambridge University Press, Cambridge, pp. xv–xxii https://doi.org/10.1017/CBO9780511809071.001.

Maragkaki, A.E., Fountoulakis, M., Kyriakou, A., Lasaridi, K., Manios, T., 2018. Boosting biogas production from sewage sludge by adding small amount of agro-industrial by-products and food waste residues. Waste Manag. 71, 605–611. https://doi.org/10.1016/j.wasman.2017.04.024.

Matuszewska, A., Owczuk, M., Zamojska-Jaroszewicz, A., Jakubiak-Lasocka, J., Lasocki, J., Orliński, P., 2016. Evaluation of the biological methane potential of various feedstock for the production of biogas to supply agricultural tractors. Energy Convers. Manag. 125, 309–319. https://doi.org/10.1016/j.enconman.2016.02.072.

Menon, A., Wang, J.-Y., Giannis, A., 2017. Optimization of micronutrient supplement for enhancing biogas production from food waste in two-phase thermophilic anaerobic digestion. Waste Manag. 59, 465–475. https://doi.org/10.1016/j.wasman.2016.10.017.

Momtazi, S., 2018. Unsupervised Latent Dirichlet Allocation for supervised question classification. Inf. Process. Manag. 54, 380–393. https://doi.org/10.1016/j.ipm.2018.01.001.

MosayebNezhad, M., Mehr, A.S., Lanzini, A., Misul, D., Santarelli, M., 2019. Technology review and thermodynamic performance study of a biogas-fed micro humid air turbine. Renew. Energy 140, 407–418. https://doi.org/10.1016/j.renene.2019.03.064.

Nabli, H., Ben Djemaa, R., Ben Amor, I.A., 2018. Efficient cloud service discovery approach based on LDA topic modeling. J. Syst. Softw. 146, 233–248. https://doi.org/10.1016/j.jss.2018.09.069.

Negro, V., Mancini, G., Ruggeri, B., Fino, D., 2016. Citrus waste as feedstock for bio-based products recovery: review on limonene case study and energy valorization. Bioresour. Technol. 214, 806–815. https://doi.org/10.1016/j.biortech.2016.05.006.

Nguyen, D.D., Yeop, J.S., Choi, J., Kim, S., Chang, S.W., Jeon, B.-H., Guo, W., Ngo, H.H., 2017. A new approach for concurrently improving performance of South Korean food waste valorization and renewable energy recovery via dry anaerobic digestion under mesophilic and thermophilic conditions. Waste Manag. 66, 161–168. https://doi.org/10.1016/j.wasman.2017.03.049.

Nie, Y., Tian, X., Zhou, Z., Li, Y.-Y., 2017. Impact of food to microorganism ratio and alcohol ethoxylate dosage on methane production in treatment of low-strength wastewater by a submerged anaerobic membrane bioreactor. Front. Environ. Sci. Eng. 11 (6). https://doi.org/10.1007/s11783-017-0947-1.

Oshita, K., Omori, K., Takaoka, M., Mizuno, T., 2014. Removal of siloxanes in sewage sludge by thermal treatment with gas stripping. Energy Convers. Manag. 81, 290–297. https://doi.org/10.1016/j.enconman.2014.02.050.

Pavlinek, M., Podgorelec, V., 2017. Text classification method based on self-training and LDA topic models. Expert Syst. Appl. 80, 83–93. https://doi.org/10.1016/j.eswa.2017.03.020.

Pérez, J., Pérez, A., Casillas, A., Gojenola, K., 2018. Cardiology record multi-label classification using latent Dirichlet allocation. Comput. Methods Prog. Biomed. 164, 111–119. https://doi.org/10.1016/j.cmpb.2018.07.002.

Pourbafrani, M., Forgács, G., Horváth, I.S., Niklasson, C., Taherzadeh, M.J., 2010. Production of biofuels, limonene and pectin from citrus wastes. Bioresour. Technol. 101, 4246–4250. https://doi.org/10.1016/j.biortech.2010.01.077.

Rajagopal, R., Bellavance, D., Rahaman, M.S., 2017. Psychrophilic anaerobic digestion of semi-dry mixed municipal food waste: for North American context. Process. Saf. Environ. Prot. 105, 101–108. https://doi.org/10.1016/j.psep.2016.10.014.

Reddy, K.S., Aravindhan, S., Mallick, T.K., 2016. Investigation of performance and emission characteristics of a biogas fuelled electric generator integrated with solar concentrated photovoltaic system. Renew. Energy 92, 233–243. https://doi.org/10.1016/j.renene.2016.02.008.

Ren, Y., Yu, M., Wu, C., Wang, Q., Gao, M., Huang, Q., Liu, Y., 2018. A comprehensive review on food waste anaerobic digestion: research updates and tendencies. Bioresour. Technol. 247, 1069–1076. https://doi.org/10.1016/j.biortech.2017.09.109.

Sahota, S., Shah, G., Ghosh, P., Kapoor, R., Sengupta, S., Singh, P., Vijay, V., Sahay, A., Vijay, V.K., Thakur, I.S., 2018. Review of trends in biogas upgradation technologies and future perspectives. Bioresour. Technol. Reports 1, 79–88. https://doi.org/10.1016/j.biteb.2018.01.002.

Silge, J., Robinson, D., 2017. Text Mining With R: A Tidy Approach. O'Reilly Media.

Song, K., Kim, K., Lee, S., 2018. Identifying promising technologies using patents: a retrospective feature analysis and a prospective needs analysis on outlier patents. Technol. Forecast. Soc. Change 128, 118–132. https://doi.org/10.1016/j.techfore.2017.11.008.

Souili, A., Cavallucci, D., Rousselot, F., 2015. A lexico-syntactic pattern matching method to extract Idm- Triz knowledge from on-line patent databases. Procedia Eng 131, 418–425. https://doi.org/10.1016/j.proeng.2015.12.437.

Thyberg, K.L., Tonjes, D.J., 2017. The environmental impacts of alternative food waste treatment technologies in the U.S. J. Clean. Prod. 158, 101–108. https://doi.org/10.1016/j.jclepro.2017.04.169.

Tong, H., Shen, Y., Zhang, J., Wang, C.-H., Ge, T.S., Tong, Y.W., 2018. A comparative life cycle assessment on four waste-to-energy scenarios for food waste generated in eateries. Appl. Energy 225, 1143–1157. https://doi.org/10.1016/j.apenergy.2018.05.062.

Trstenjak, B., Mikac, S., Donko, D., 2014. KNN with TF-IDF based framework for text categorization. Procedia Eng 69, 1356–1364. https://doi.org/10.1016/j.proeng.2014.03.129.

Venugopalan, S., Rai, V., 2015. Topic based classification and pattern identification in patents. Technol. Forecast. Soc. Change 94, 236–250. https://doi.org/10.1016/j.techfore.2014.10.006.

Wang, W., Feng, Y., Dai, W., 2018. Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. Electron. Commer. Res. Appl. 29, 142–156. https://doi.org/10.1016/j.elerap.2018.04.003.

Wang, Y., Xu, W., 2018. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decis. Support. Syst. 105, 87–95. https://doi.org/10.1016/j.dss.2017.11.001.

Wongso, R., Luwinda, F.A., Trisnajaya, B.C., Rusli, O., Rudy, A., 2017. News article text classification in Indonesian language. Procedia Comput. Sci. 116, 137–143. https://doi.org/10.1016/j.procs.2017.10.039.

Woon, K.S., Lo, I.M.C., Chiu, S.L.H., Yan, D.Y.S., 2016. Environmental assessment of food waste valorization in producing biogas for various types of energy use based on LCA approach. Waste Manag. 50, 290–299. https://doi.org/10.1016/j.wasman.2016.02.022.

Wu, J.-L., Chang, P.-C., Tsao, C.-C., Fan, C.-Y., 2016. A patent quality analysis and classification system using self-organizing maps with support vector machine. Appl. Soft Comput. 41, 305–316. https://doi.org/10.1016/j.asoc.2016.01.020.

Wyman, V., Henríquez, J., Palma, C., Carvajal, A., 2018. Lignocellulosic waste valorisation strategy through enzyme and biogas production. Bioresour. Technol. 247, 402–411. https://doi.org/10.1016/j.biortech.2017.09.055.

Ye, M., Liu, J., Ma, C., Li, Y.-Y., Zou, L., Qian, G., Xu, Z.P., 2018. Improving the stability and efficiency of anaerobic digestion of food waste using additives: a critical review. J. Clean. Prod. 192, 316–326. https://doi.org/10.1016/j.jclepro.2018.04.244.

Zhang, J., Li, W., Lee, J., Loh, K.-C., Dai, Y., Tong, Y.W., 2017a. Enhancement of biogas production in anaerobic co-digestion of food waste and waste activated sludge by biological co-pretreatment. Energy 137, 479–486. https://doi.org/10.1016/j.energy.2017.02.163.

Zhang, J., Loh, K.-C., Li, W., Lim, J.W., Dai, Y., Tong, Y.W., 2017b. Three-stage anaerobic digester for food waste. Appl. Energy 194, 287–295. https://doi.org/10.1016/j.apenergy.2016.10.116.

Zhang, P., Gu, H., Gartrell, M., Lu, T., Yang, D., Ding, X., Gu, N., 2016. Group-based Latent Dirichlet Allocation (Group-LDA): effective audience detection for books in online social media. Knowledge-Based Syst 105, 134–146. https://doi.org/10.1016/j.knosys.2016.05.006.

Zhang, W., Lang, Q., Fang, M., Li, X., Bah, H., Dong, H., Dong, R., 2017. Combined effect of crude fat content and initial substrate concentration on batch anaerobic digestion characteristics of food waste. Bioresour. Technol. 232, 304–312. https://doi.org/10.1016/j.biortech.2017.02.039.

Zhang, X., 2014. Interactive patent classification based on multi-classifier fusion and active learning. Neurocomputing 127, 200–205. https://doi.org/10.1016/j.neucom.2013.08.013.

Zhou, Z., Chen, L., Wu, Q., Zheng, T., Yuan, H., Peng, N., He, M., 2019. The valorization of biogas slurry with a pilot dual stage reverse osmosis membrane process. Chem. Eng. Res. Des. 142, 133–142. https://doi.org/10.1016/j.cherd.2018.12.005.