Innovation hotspots in food waste treatment, biogas, and anaerobic digestion technology: a natural language processing approach on patent data

Djavan De Clercq[a,b], Zongguo Wen[a,b,*], Jade De Clercq[c]

a. State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, 100084 China
b. Key Laboratory for Solid Waste Management and Environment Safety (Tsinghua University), Ministry of Education of China, Tsinghua University, Beijing 100084, China
c. Department of Materials, Imperial College London, United Kingdom

* Corresponding author: wenzg@tsinghua.edu.cn

Abstract

The objective of this study is to explore the current landscape and emerging trends in innovation related to food waste treatment, biogas, and anaerobic digestion. The methodology used involved analyzing large volumes of text data mined from 3,186 patents related to these three fields. Latent Dirichlet Allocation and the perplexity method were used to identify the main topics which these patent corpora were comprised of and which technological concepts were most associated with each topic. In addition, 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.

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 regards to technology selection are proposed based on the analysis.
- The entire data and code behind the analysis is open-source and available online.

Keywords

- Natural language processing; Latent Dirichlet Allocation; TF-IDF; food waste; biogas, anaerobic digestion.

1. Introduction

1.1. Anaerobic food waste technology roadmapping is difficult due to a plethora of conflicting information

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 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 are 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 (Algapani et al., 2019; 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; Qin et al., 2018; Rajagopal et al., 2017; Tonanzi et al., 2018; Ye et al., 2018; J. 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. Several other recent studies have applied LCA to environmental impact evaluation of food waste treatment technology (Cristóbal et al., 2016; Edwards et al., 2017; Jin et al., 2015; Lijó et al., 2017; Woon et al., 2016; Xu et al., 2015).

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 useful information for technology policy roadmapping

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 tools such as natural language processing 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 (Joung and Kim, 2017; Song et al., 2018; Suominen et al., 2017).

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. Venugopalan and Rai (2015) built a classifier based on document-term frequency and topic modelling in order to categorize 10,201 patents about solar photovoltaics by technology area.

1.3. Natural language processing techniques have not been applied to food waste, AD, and biogas patents

NLP 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.

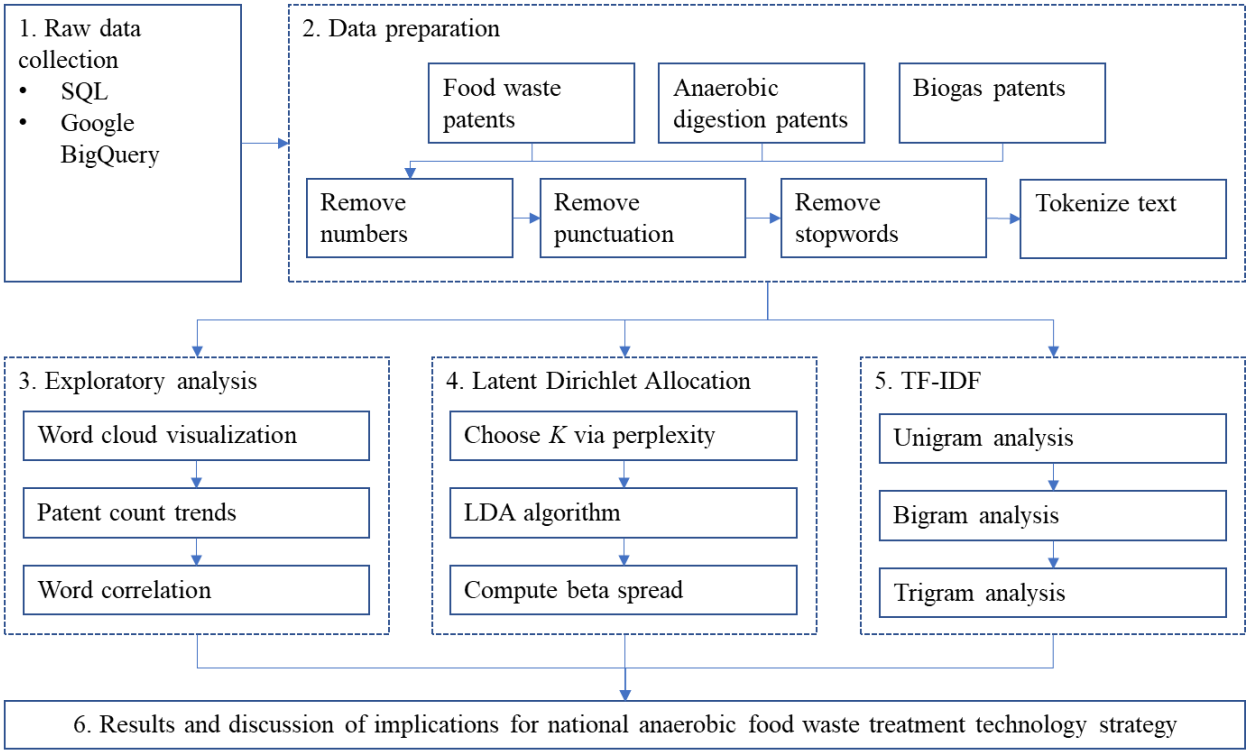
The study addresses the following research question: what are the trends in patenting activity related to food waste, anaerobic digestion, and biogas? To answer this question, we apply Latent Dirichlet Allocation and TF-IDF to (1) identifying the topics covered in these patents and (2) identifying emerging 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). Although

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 in section 5 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 figure 1. Firstly, raw data was collected based on SQL query of an open-source 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.

Figure 1: Overview of methodology



Each sub-section below corresponds to the numbered box in figure 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. 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, Japane, 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.

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 Robinson, 2017). Given table 1:

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

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); 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 ($z_1, z_2, \dots, z_k$), 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 (Li et al., 2018):

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_K^{\alpha_K - 1}, \tag{2}$$

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):

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

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

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

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

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

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

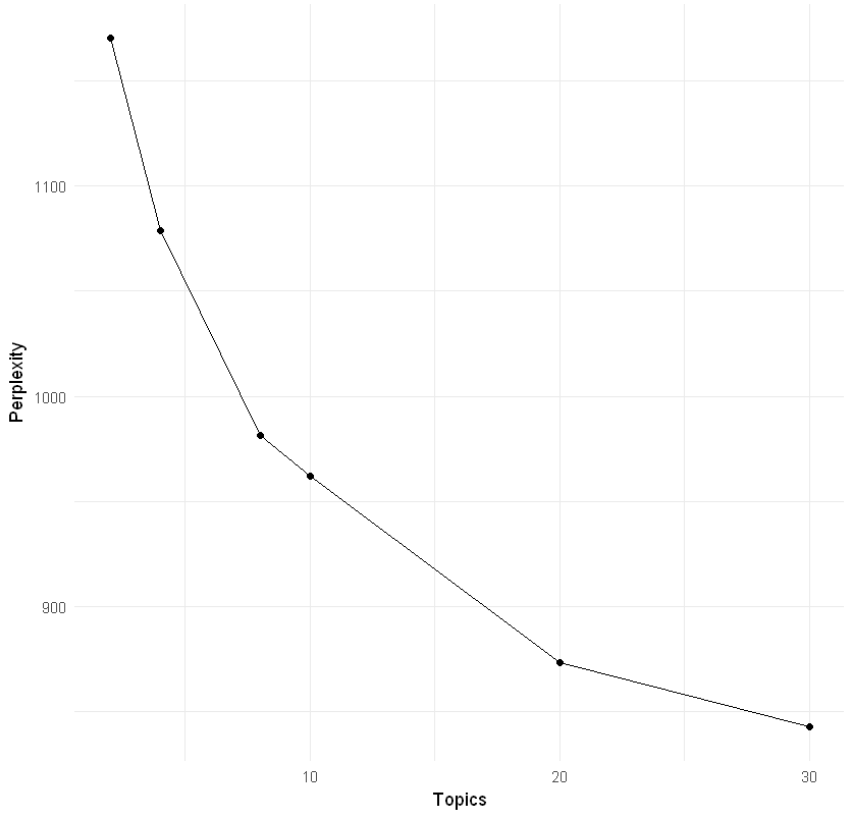
In the context of this study, the objective of LDA is to find the mixture of words that is associated with each of the K topics, and to determine the mixture of topics that describes each document. Patents are then labelled with K topic probabilities, indicating the likelihood that a given patent is 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). Perplexity can be calculated with the following equation (Pavlinek and Podgorelec, 2017):

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

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.

Figure 2: Perplexity on the validation set depending on the number of topics



Lower values of perplexity indicate lower misrepresentation of the words of the test documents by the trained topics. Figure 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

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):

$$\mathrm{idf}_t = \log\left(\frac{N}{\mathrm{df_t}}\right) \tag{7}$$

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* given by

$$\mathrm{tf\_idf}_{t,d} = \mathrm{tf}_{t,d} * \mathrm{idf_t} \tag{8}$$

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 was used as an IDE, and the entire notebook code behind this study is open-source and can be found online: XXXX（还要把 data 与代码上传）

3. Results & Discussion

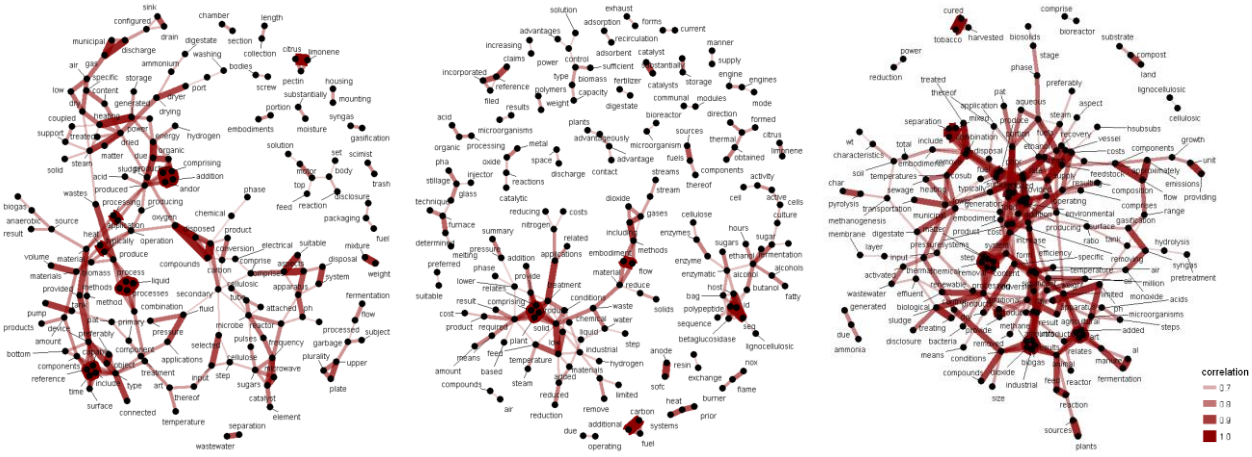
3.1. Exploratory analysis

S1 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 figure 3 demonstrate which keywords occur more often together than other keywords. Notice the presence of both larger, interconnected clusters and smaller, distinct clusters.

Figure 3: pairs of words in the patent corpora that show at least a 0.60 correlation of appearing within the same 4-line section



Food waste Biogas Anaerobic digestion

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 via an integrated process, finding that one ton of citrus waste resulted in 39.64 liters of ethanol, 45 m^3 of biogas, 8.9 liters of limonene, and 38.8 kg of pectin. Other distinct clusters included: (1) “separation” and “wastewater”; (2) “chamber” and “section”; (3) “syngas” and “gasification”.

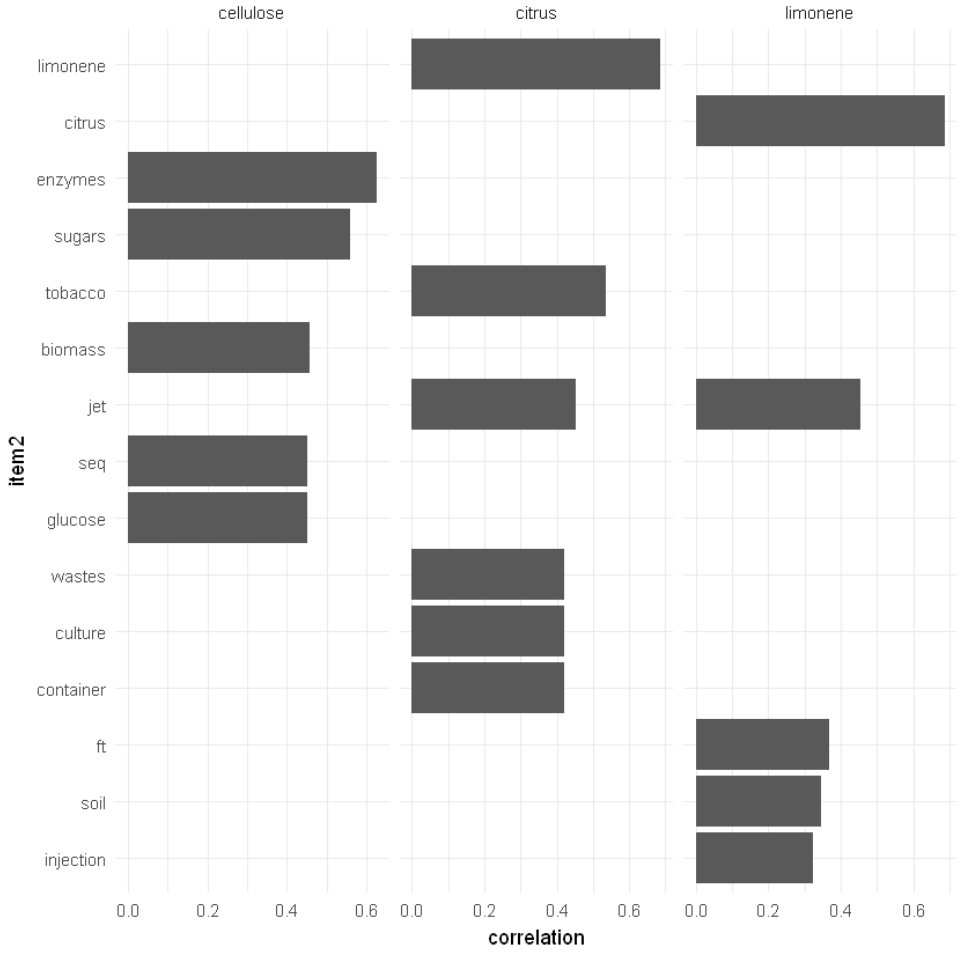
In the biogas patent corpus, the correlation network revealed even more distinct clusters of words, possibly due to the more limited technical scope of biogas compared to food waste. As in the food waste corpus, the words “citrus” and “limonene” formed a distinct cluster. Other notable clusters included: (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”.

In addition to the networks visualized in figure 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, figure 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”.

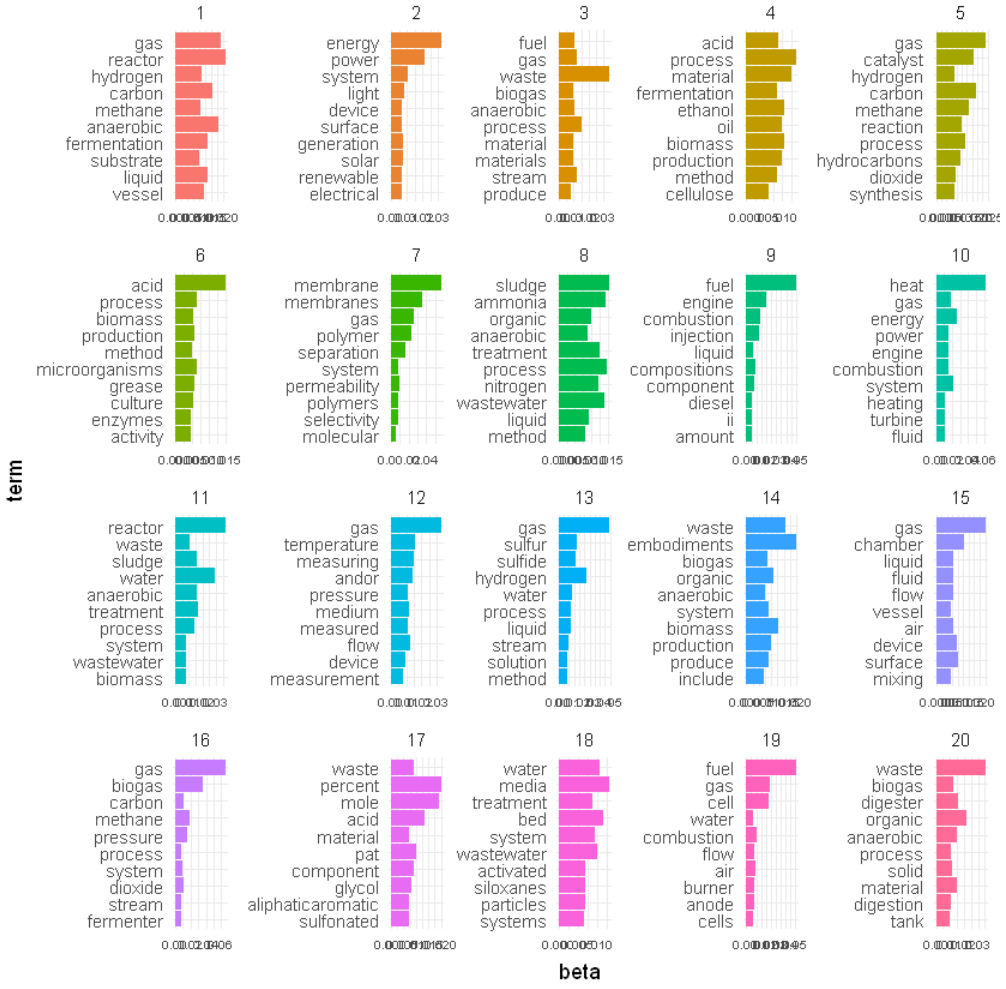
Figure 4: The top five words from the biogas patent corpus that were most correlated with “cellulose”, “citrus”, and “limonene” in a 4-line section



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 were computed. Figure 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.

Figure 5: Top 10 words in each LDA topics for the biogas patent corpus



For instance, figure 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:

$$\log_2\left(\frac{\beta_2}{\beta_1}\right) \tag{8}$$

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.

Figure 6: words with the greatest difference in β between topic 1 and topic 2 (biogas patent corpus)

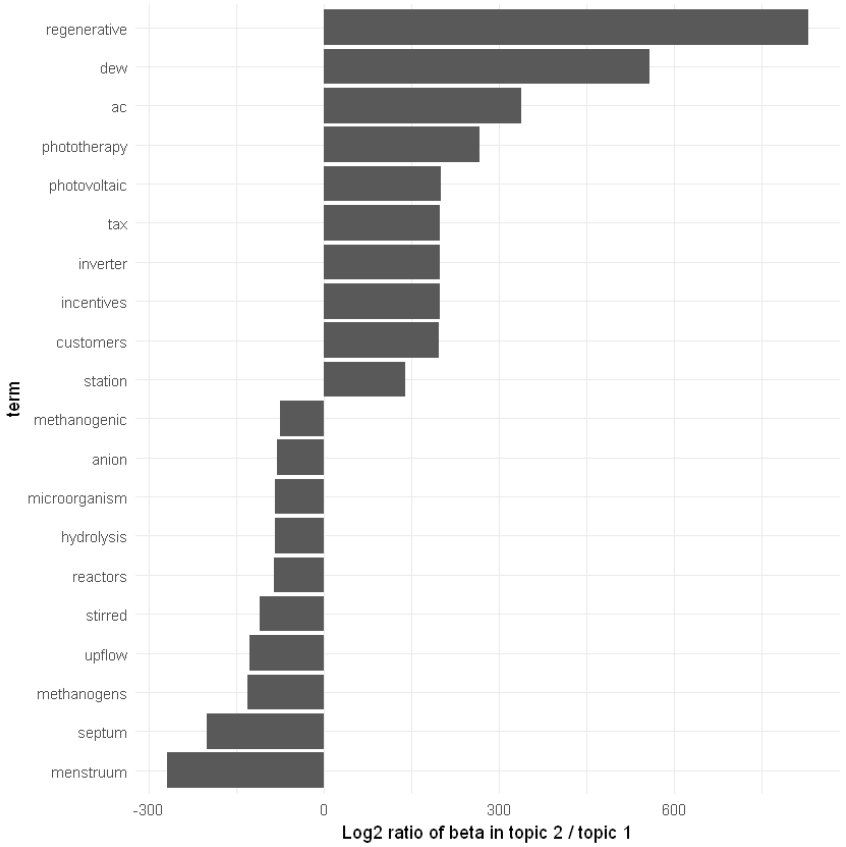
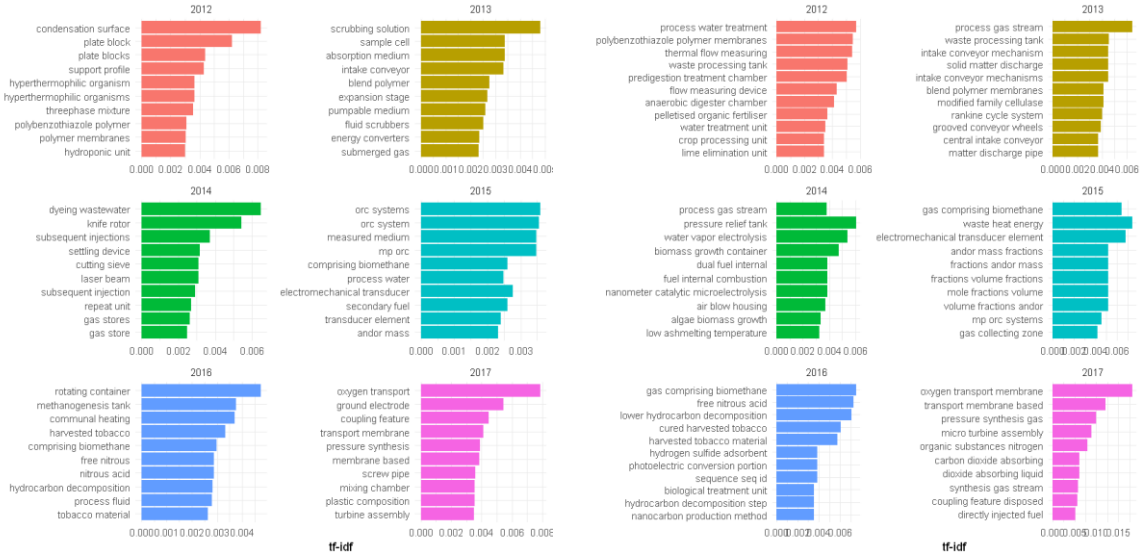


Figure 6 provides a good example of this, and shows that discriminatory words for topic 1 include “menstruum”, “septum”, “methanogens”, “upflow”, “stirred”, and “reactor”, indicating that topic 1 is closely related to specifics of gas production in a digester. Topic 2 words include “regenerative”, “dew”, “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.

3.5. 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. Figure 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 .

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



The left panel of figure 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”.

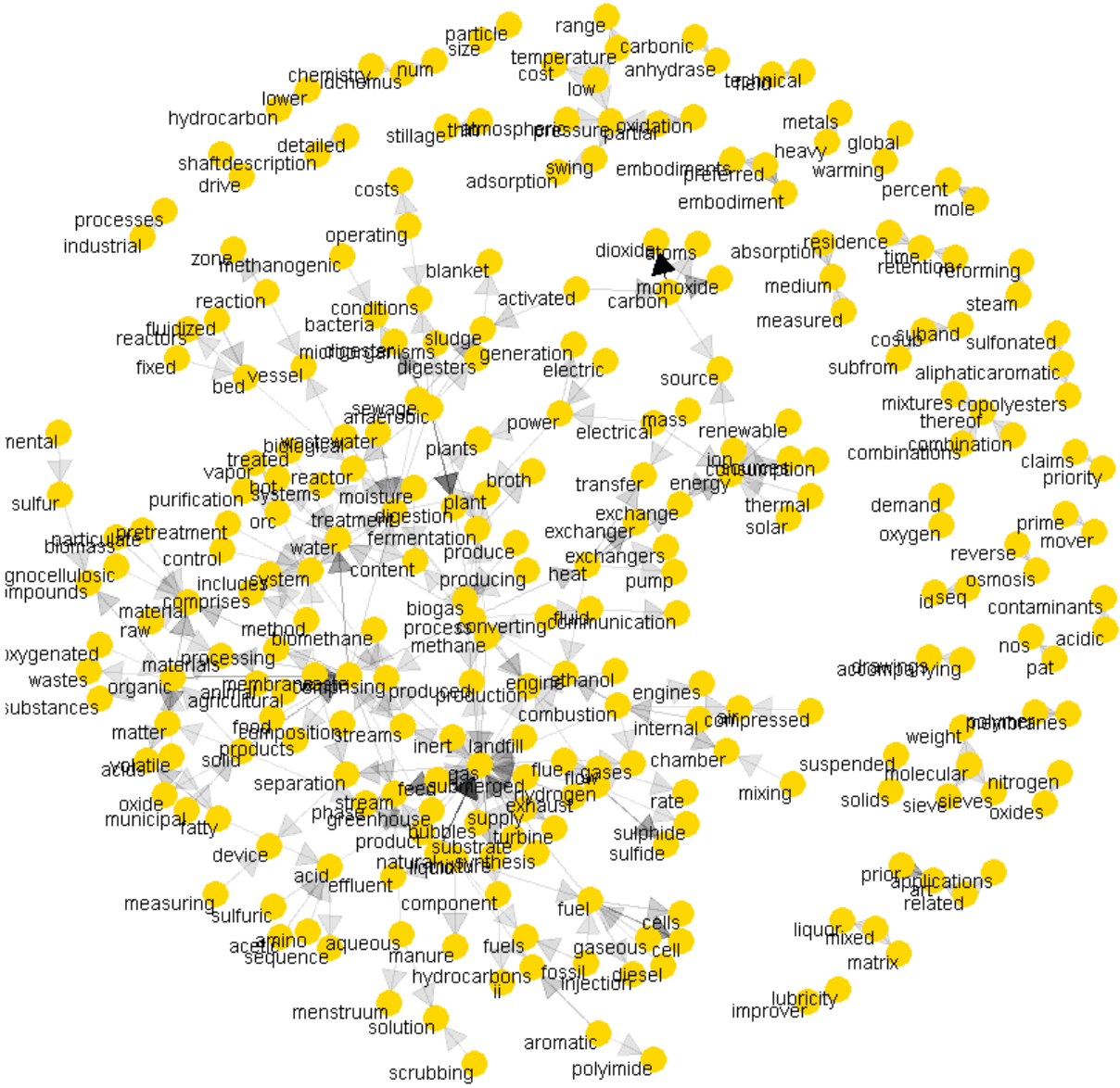
The right panel of figure 7 offers additional granularity by showing trigrams (three words) with the highest TF-IDF in the biogas corpus across several years. For instance, in 2017, they reveal patenting related to concepts including “oxygen transport membrane”, “pressure synthesis gas”, “micro turbine 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.

In addition to visualizing the top few n-grams as shown in figure 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 figure 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 figure 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 isolation of technology concepts that representative of this particular body of patent text. 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.

Figure 8: Common bigrams the biogas patent corpus that occurred at least 100 times



4. Conclusion and policy suggestions

This study applied natural language processing methods to (1) identifying the technological concepts (topics) present in patent documents related to food waste treatment, biogas, and anaerobic digestion. The entire data set and R code behind this study can be found in an online supplement.

Latent Dirichlet Allocation was conducted on the text data to identify the main topics which describe the patent corpora. In addition, TF-IDF was applied to characterizing emerging technology trends in the patent corpora across selected years.

Some selected results of the LDA analysis include the following:

- The perplexity measure indicated that 20-topic models were appropriate for the three sets of patent documents.
- LDA topics showed 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”.

Some selected 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”.

Governments 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.

Acknowledgements

The authors gratefully acknowledge the financial support from the "Thirteenth Five-Year" National Key R&D Program of China (2016YFC0502800) and General Programs of the National Natural Science Foundation of China (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.

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/https://doi.org/10.1016/j.actaastro.2017.06.029

Algapani, D.E., Qiao, W., Ricci, M., Bianchi, D., M. Wandera, S., Adani, F., Dong, R., 2019. Bio-hydrogen and bio-methane production from food waste in a two-stage anaerobic digestion process with digestate recirculation. Renew. Energy 130, 1108–1115. https://doi.org/https://doi.org/10.1016/j.renene.2018.08.079

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/https://doi.org/10.1016/j.jclepro.2016.12.049

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/https://doi.org/10.1016/j.jclepro.2018.07.240

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/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/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. Sustain. Energy Rev. 79, 204–221. https://doi.org/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/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/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/https://doi.org/10.1016/j.biortech.2016.10.044

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/https://doi.org/10.1016/j.egypro.2017.03.811

Jin, Y., Chen, T., Chen, X., Yu, Z., 2015. Life-cycle assessment of energy consumption and environmental impact of an integrated food waste-based biogas plant. Appl. Energy 151, 227–236. https://doi.org/https://doi.org/10.1016/j.apenergy.2015.04.058

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/https://doi.org/10.1016/j.techfore.2016.08.020

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/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/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/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/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/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/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/https://doi.org/10.1016/j.apenergy.2017.05.042

Lijó, L., Lorenzo-Toja, Y., González-García, S., Bacenetti, J., Negri, M., Moreira, M.T., 2017. Eco-efficiency assessment of farm-scaled biogas plants. Bioresour. Technol. 237, 146–155. https://doi.org/https://doi.org/10.1016/j.biortech.2017.01.055

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/https://doi.org/10.1016/j.biortech.2016.07.130

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/https://doi.org/10.1016/j.wpi.2016.05.008

Manning, C.D., Schütze, H., Raghavan, P. (Eds.), 2008. Preface, in: Introduction to Information Retrieval. Cambridge University Press, Cambridge, pp. xv–xxii. https://doi.org/DOI: 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/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/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/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/https://doi.org/10.1016/j.ipm.2018.01.001

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/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/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/https://doi.org/10.1016/j.eswa.2017.03.020

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/https://doi.org/10.1016/j.biortech.2010.01.077

Qin, Y., Wu, J., Xiao, B., Hojo, T., Li, Y.-Y., 2018. Biogas recovery from two-phase anaerobic digestion of food waste and paper waste: Optimization of paper waste addition. Sci. Total Environ. 634, 1222–1230. https://doi.org/https://doi.org/10.1016/j.scitotenv.2018.03.341

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/https://doi.org/10.1016/j.psep.2016.10.014

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/https://doi.org/10.1016/j.biortech.2017.09.109

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/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/https://doi.org/10.1016/j.proeng.2015.12.437

Suominen, A., Toivanen, H., Seppänen, M., 2017. Firms' knowledge profiles: Mapping patent data with unsupervised learning. Technol. Forecast. Soc. Change 115, 131–142. https://doi.org/https://doi.org/10.1016/j.techfore.2016.09.028

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/https://doi.org/10.1016/j.jclepro.2017.04.169

Tonanzi, B., Gallipoli, A., Gianico, A., Montecchio, D., Pagliaccia, P., Di Carlo, M., Rossetti, S., Braguglia, C.M., 2018. Long-term anaerobic digestion of food waste at semi-pilot scale: Relationship between microbial community structure and process performances. Biomass and Bioenergy 118, 55–64. https://doi.org/https://doi.org/10.1016/j.biombioe.2018.08.001

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/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/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/https://doi.org/10.1016/j.techfore.2014.10.006

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/https://doi.org/10.1016/j.dss.2017.11.001

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/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/https://doi.org/10.1016/j.asoc.2016.01.020

Xu, C., Shi, W., Hong, J., Zhang, F., Chen, W., 2015. Life cycle assessment of food waste-based biogas generation. Renew. Sustain. Energy Rev. 49, 169–177. https://doi.org/https://doi.org/10.1016/j.rser.2015.04.164

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/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/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/https://doi.org/10.1016/j.apenergy.2016.10.116

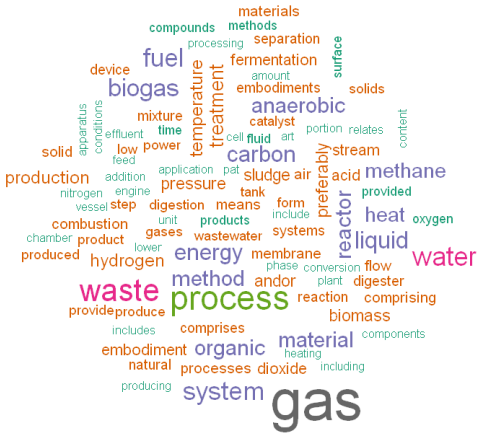
Zhang, W., Lang, Q., Fang, M., Li, X., Bah, H., Dong, H., Dong, R., 2017. Combined effect of crude fat content and initial substrate concentration on batch anaerobic digestion characteristics of food waste. Bioresour. Technol. 232, 304–312. https://doi.org/https://doi.org/10.1016/j.biortech.2017.02.039

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

Supplementary information

S.1. Word clouds showing top words in each set of patents

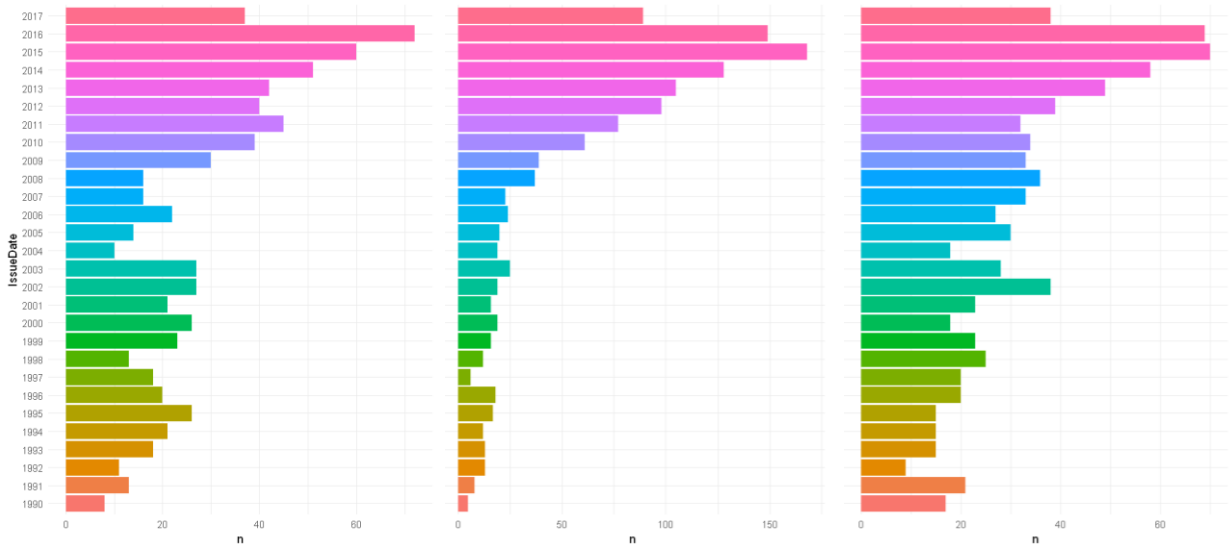
Food waste



Biogas

Anaerobic digestion

S2. Patenting activity in each patent corpus, 1990 – 2017

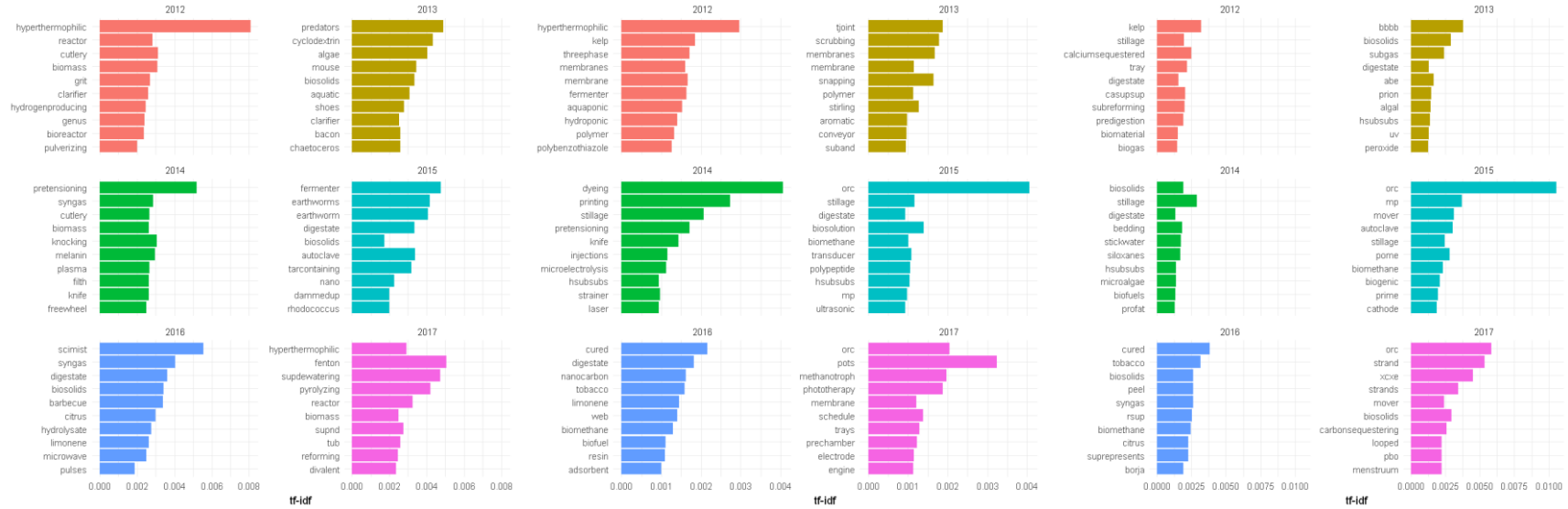


Food waste

Biogas

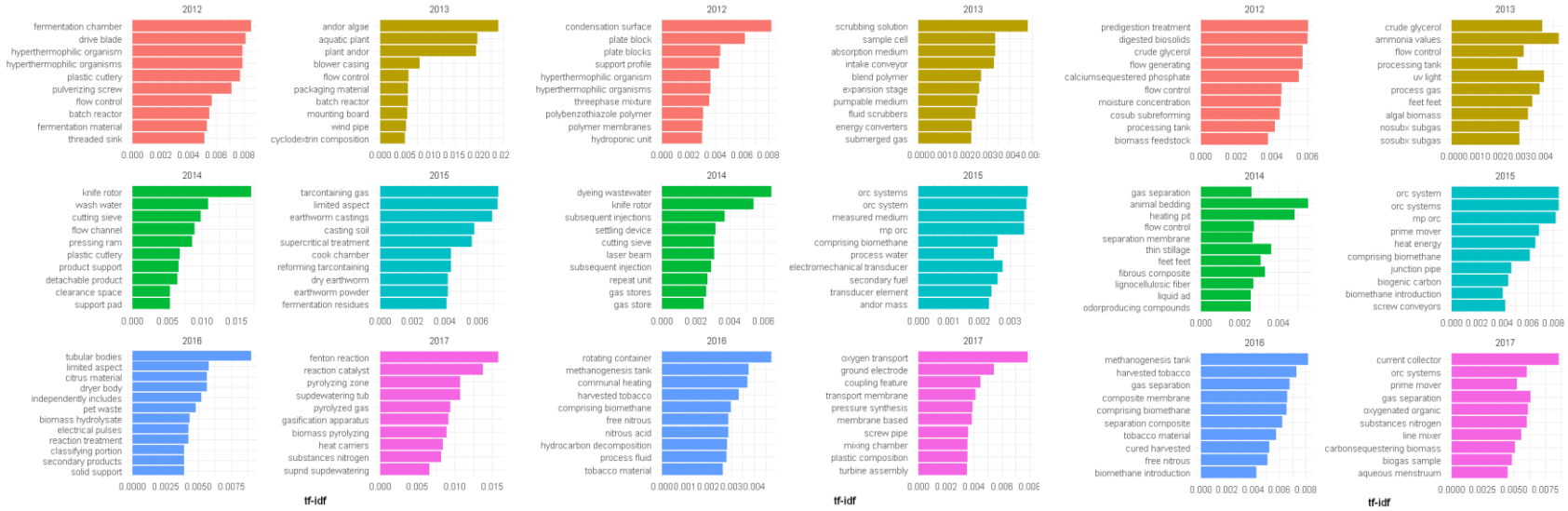
Anaerobic digestion

S3. TF-IDF analysis: unigrams (2012 – 2017)



Food waste Biogas Anaerobic digestion

S4. TF-IDF analysis: bigrams (2012 – 2017)

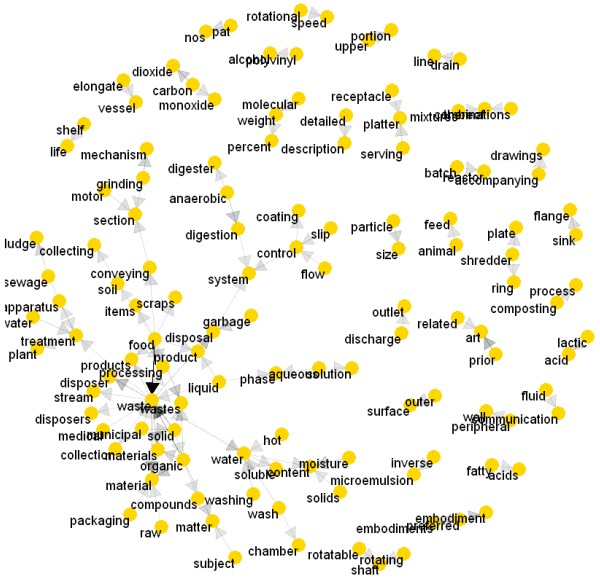


Food waste Biogas Anaerobic digestion

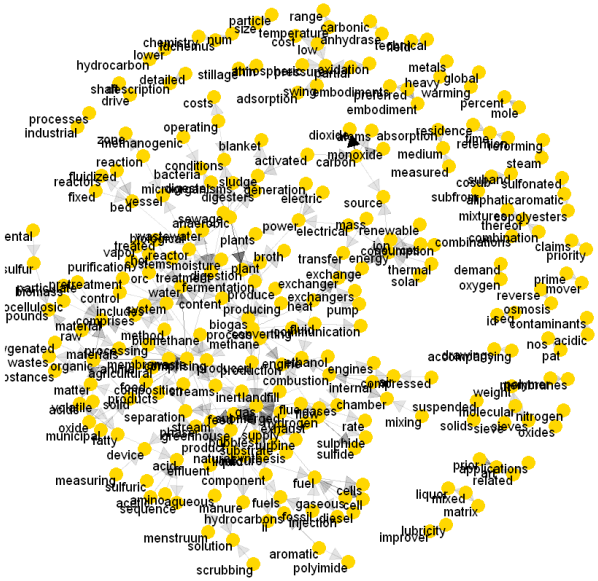
S5. TF-IDF analysis: trigrams (2012 – 2017)

Food waste Biogas Anaerobic digestion

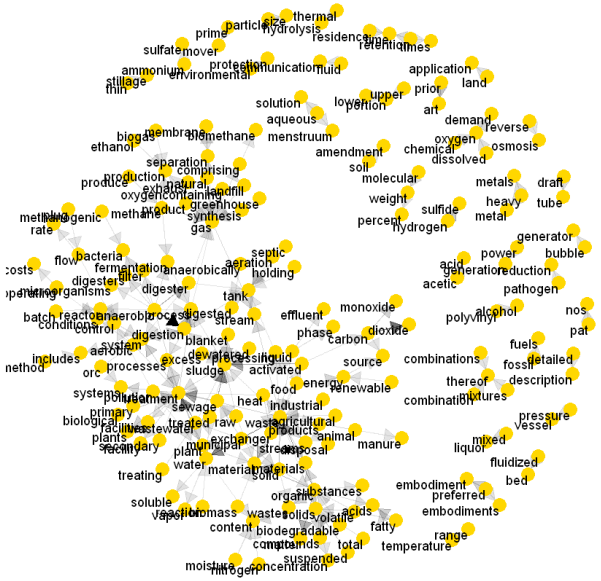
S6. Bigram network



Food waste

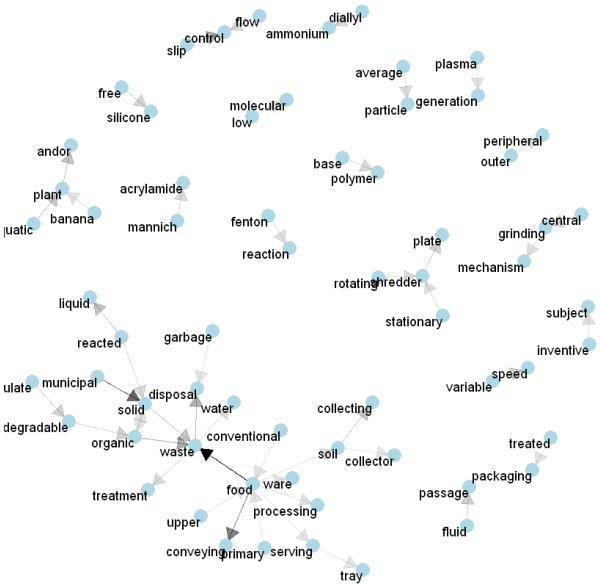


Biogas

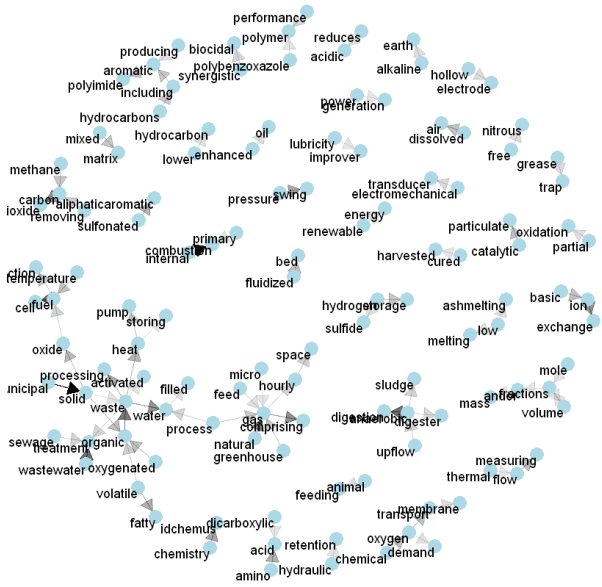


Anaerobic digestion

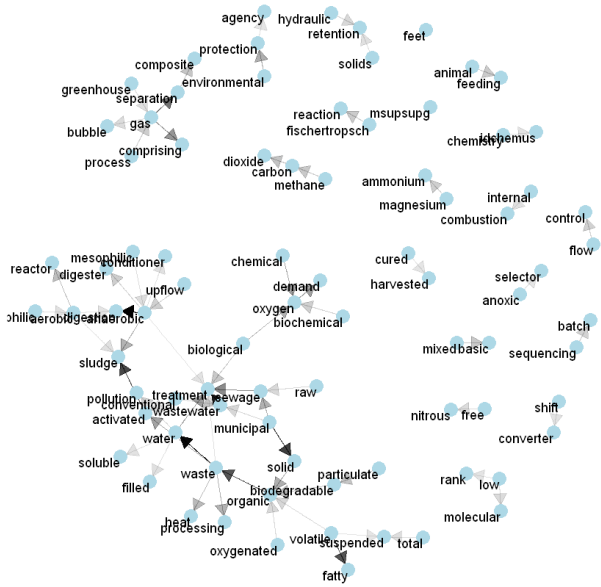
S7. Trigram network



Food waste

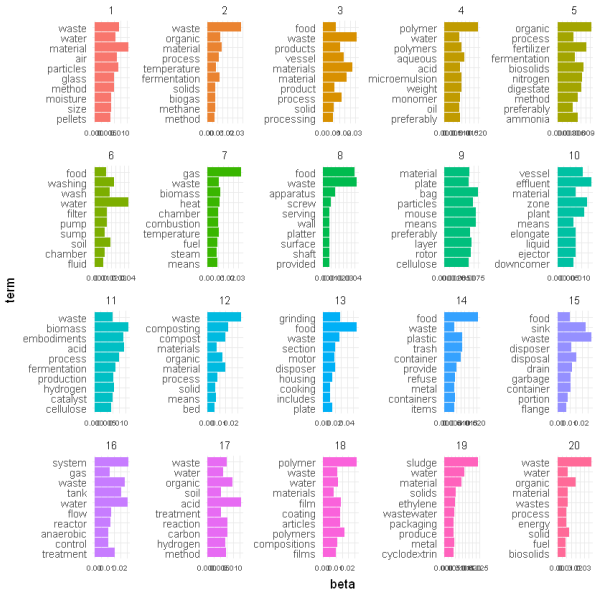


Biogas

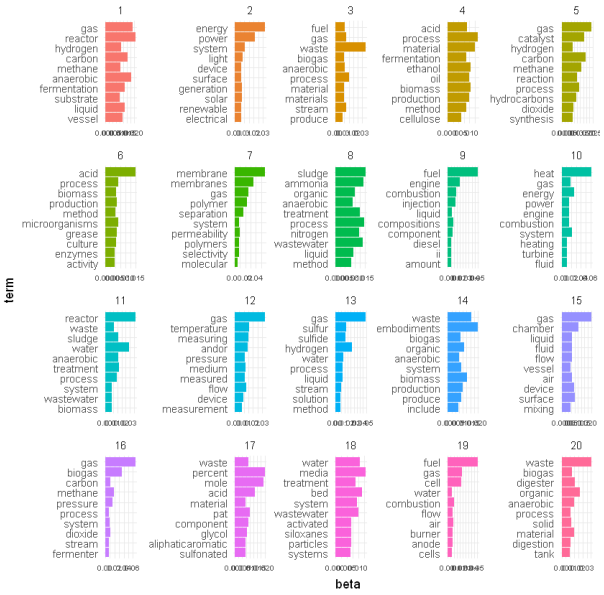


Anaerobic digestion

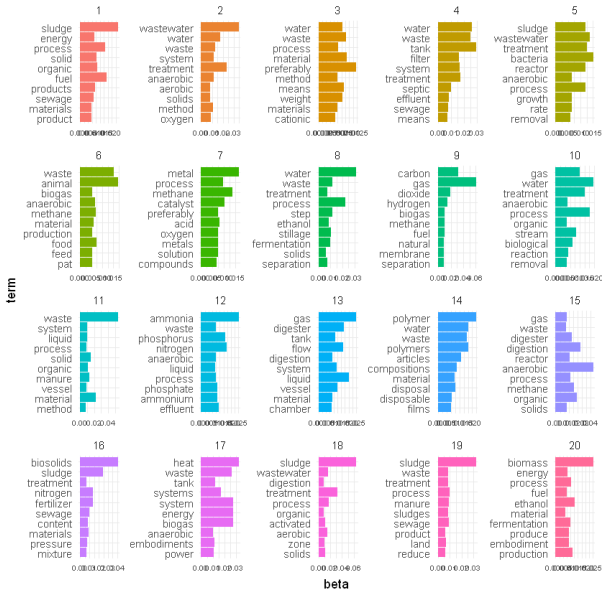
S8. LDA: words associated with each of the 20 topics for each patent corpus set



Food waste

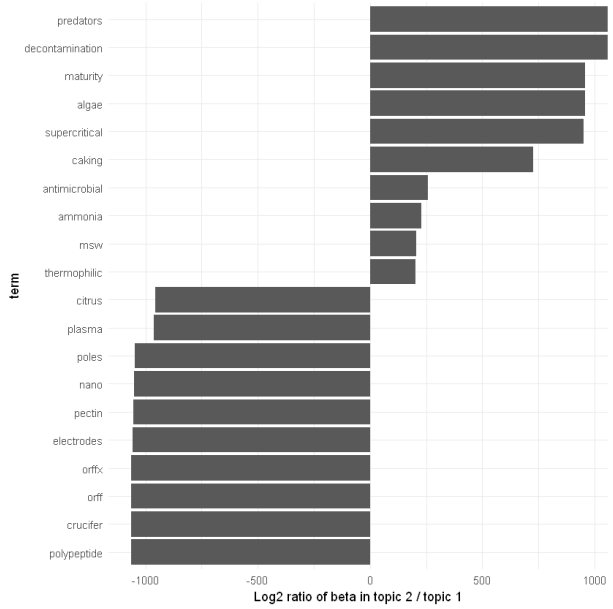


Biogas

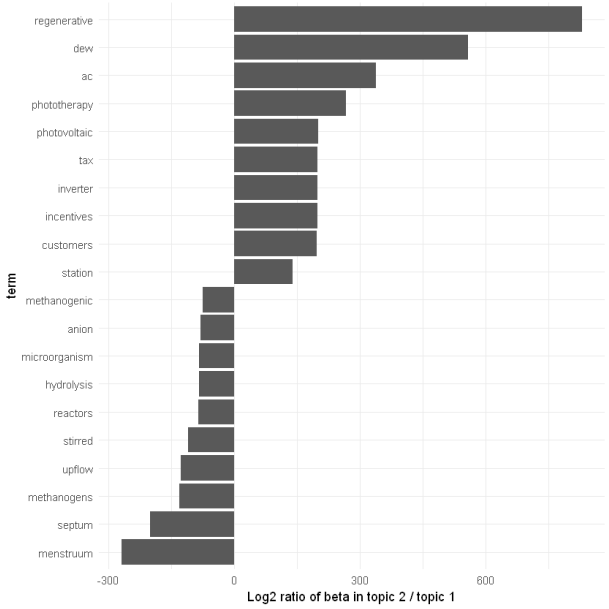


Anaerobic digestion

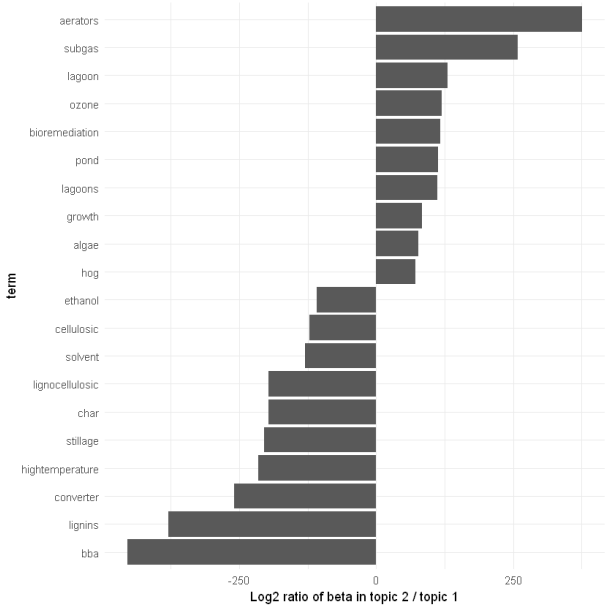
S9. Beta spread between two topics



Food waste



Biogas



Anaerobic digestion