***Science* Manuscript Template Instructions**

**General instructions on using this template and submitting a manuscript to *Science*:** Thank you for preparing a manuscript for submission to *Science*. Using this template, or following the guidelines below, will help us in processing your paper and completing the manuscript’s record in our system. Our goal is to automatically identify each section of your manuscript so that we can accurately extract title, authors, abstract, etc. and to enrich it by including reference links and an accurate layout.

Please use the actual template, which starts on page 2. When you are ready to submit, please delete the text on this cover page.

You can submit your paper at <https://cts.sciencemag.org>. Additional information for authors is available at [http://www.sciencemag.org/authors/science-information-authors](http://www.sciencemag.org/site/feature/contribinfo/index.xhtml).

If you are using LaTeX, please convert your paper into a Word .docx file. If this is not possible, please use our LaTeX template (available [here](http://www.sciencemag.org/authors/preparing-manuscripts-using-latex)) and upload a PDF version of your paper. Some conversion approaches are available here: <http://www.tug.org/utilities/texconv/textopc.html>.

So that we can extract parts of your paper (even if you do not use this template), begin each section with the specific words listed below, some of which are followed by a colon. Do not use paragraph breaks in the title, author list, or abstract. The author list, corresponding author email(s), and affiliation(s) should be checked carefully because they will be published as listed in the manuscript.

**Title:** No more than 96 characters and spaces, lacking jargon and abbreviations where possible.

**Authors:**

**Affiliations:**

**Abstract:** 125 words or less.

**One Sentence Summary:** No more than 125 characters and spaces.

**Main Text:**

**References and Notes:** Only a single numbered list should be provided for all references cited in the main text and in the supplementary materials.

**Acknowledgments:** Split into general, Funding, Author contributions, Competing interests, and Data and materials availability, as described in the template below.

**Supplementary Materials:** Include a list, noting which references are only cited in the SM.

**Fig. #.** (Begin each figure caption with a label, “**Fig. 1.**”, for example, as a new paragraph.)

**Table #.** (Begin each table caption with a label “**Table 1.**”, for example, as a new paragraph.)

Please use the .docx format (all versions after Word 2007 for PC and Word 2011 for Mac) and include page numbers in your submitted file. We also encourage use of line numbers. Supplementary Materials (comprising Materials and Methods, figures, and tables) should be in a separate file.

More specific formatting instructions are provided in the template that follows.

Title: Using Text Mining and Topic Modelling to understand the decision criteria in Renewable Energy Project Sanctioning.

**Authors:** Munish Kumar1,2 and Janine Ng2

1ERCE Singapore, 2Singapore University of Social Sciences

\*Corresponding author. Email: mkumar@erce.energy

**Abstract:** Renewable energy like solar, wind, geothermal and hydrogen are means by which corporations and government hope to mitigate the effect of climate change. Thus, it is hardly surprising that the renewable energy sector has experienced exponential growth in recent decades. This paper aims to discuss metrics that go into renewable project success, but will do so using text mining and topic modeling as a means to identify factors that contributed to renewable energy projects being sanctioned, despite difficulties like energy transportation and subpar performance of certain renewable energy plants. Text mining is applied on a collection of scientific research data, by targeting published scientific papers with “renewable energy success” in its subject matter

**One-Sentence Summary:** If renewable energy projects are non-commercial, can they be labelled as successes?

As climate concerns increase, corporations and governments are seen to make increasing investments in renewable energy (RE) technologies in such diverse areas as solar, wind, hydrogen and biomass. The expectation with such renewable technology is that harmful greenhouses gases are reduced, along with other forms of pollution. There is also a secondary consideration driven more by an economic mindset; as fossil fuels become increasing scarce, energy producers are only able to meet the growing demand at an ever-increasing cost. In either scenario however, an increase in contribution of renewable resources into the energy mix is a positive step.

However, several critical challenges still exist, and which reduces the efficacy of widespread RE adoption. First, RE tends to be a less reliable form of energy, due to the unpredictability associated with the natural environment. For instance, it was reported that in India, despite its RE footprint growing by ~20% from 2015-2020, ~15-20% of (wind and solar) projects underperformed during 2019-2020, primarily due to adverse weather conditions [1]. A mitigant to this is to use technology like batteries, but this adds capital and operational expenditures (CAPEX/ OPEX) to projects that are sometimes already financially stressed. Another challenge associated with RE adoption has to do with its effect on legacy electrical grids; without a proper understanding of import capacity limits and load factors, RE can cause a deterioration of power factors and in fact escalate costs for operators [2].

The above are just snippets of issues associated with RE, its challenging path to commerciality and why widespread adoption is sometimes problematic. Despite this, investment in RE continues to build. (Emelyn/Inggrid to help with links). This is indeed puzzling. Since RE cannot be easily/cheaply stored, transported, has sub optimal production (at times) and is expensive to build, maintain and implement, what are the key motivating factors that are driving private and public sector stakeholders to sanction or fund RE projects? Literature itself is extremely polarized on this, with some authors arguing that economic, financial and technical metrics are key, [3, 4, 5], while others argue that political will, behavioral and social aspects are sufficient, if not more important [6, 7, 8, 9].

**Scope & Methods:** This paper will attempt to understand and rank key determinants in RE project sanctioning. Given both the polarization of opinion as well as sheer volume of discourse in academic literature in this area, this question lends itself quite readily to methods in data science related to ‘Text Mining’ and ‘Topic Modeling’ (TM+TM) [10, 11].

Essentially, both techniques, part of the Natural Language Processing (NLP) set of tools, will be used to understand the patterns present in unstructured data sources like words, phrases, sentences, and strings of text. NLP algorithms transform unstructured text formats into structured data, enabling unsupervised machine learning processes to be applied. The algorithms automate procedures of categorising, clustering, tagging, and classifying texts, and can extract information on sentiment, topics or intent, all with the goal to uncover hidden structures or commonalities binding the text [10].

For this paper, NLP is implemented using Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NNMF). LDA uses statistics to identify hidden structures within text, while NNMF utilizes linear algebra. By applying weightages to individual words based on their similarities, each algorithm can cluster the family of words into a topic [12, 13]. A key assumption in LDA is that each text block contains ‘*k*’ number of topics, represented by the word distribution [13].

The techniques are applied to a collated collection of 80 scientific articles with the terms “renewable energy success” present as a subject matter. The selection of scientific articles was based on the criteria that (a) each research article should explain factors of success in renewable energy projects, and (b) the understanding of success was from each article’s own definition of success. Each article was also briefly fact checked to ensure their relevancy to the topic. In order to avoid extended processing times and to ensure that longer articles (with higher word counts) do not unnecessarily skew the evaluation, only text from the abstract and conclusion sections were analysed, the assumption being that they are the most succinct and distilled form of information pertaining to the article. The working hypothesis is that, while brief on detail, these 2 text blocks still contain sufficiently insights to arrive at a meaningful conclusion.

***Data Preparation and Preliminary Exploration:*** A complete flow chart highlighting the NLP process is given in Figure 1. Note that a modified stop word dictionary was created to remove words such as “renewable”, “energy” and “success”, as these would be high frequency words affecting the resulting distribution of words. The process of “tokenization” and “lemmatization” is as done in other NLP processes [12]. The output of this process was visualized using a word cloud where the top 10 most frequent words were highlighted; indeed, a preliminary assessment of the word cloud indicated that the randomly selected articles were suitable for use in the next stage of analysis (Figure 2).

**Diagram, timeline

Description automatically generated**

Figure 1: Flow chart highlighting text processing stages applied to the scientific articles in question.

Text

Description automatically generated

Figure 2: Top 10 words with highest frequency.

**Topic Modelling with LDA:** The LDA algorithm is implemented via the use of 2 open-source python libraries, namely Gensim and Mallet (“MAchine Learning for LanguagE Tootlkit”). The former creates a unique dictionary like structure where each word is given a unique id and frequency count of occurrence and also doubles as a visualization library. Mallet is used complementarily to Gensim for coherence analysis. Coherence here is a measure of effectiveness of the topic model. A coherence score measures the degree of sematic similarity between the top scoring words in each topic and also facilitates decisions relating to the optimal value of ‘*k’* (the number of topics). As ‘k varies, so too does the coherence score, with higher coherence scores generally representing better model.

Figure 3 shows the coherence score obtained, as a function of the number of topics. The graph shows that the coherence scores do not varying much between the ranges of 5 to 7 and thereafter decrease. In order to demine topic distinctiveness and prevalence, an intertopic distance map (Figure 4, left) is used, with each bubble representing a topic. Through the use of multidimension reduction scaling (PCA/t-sne), topic’s probability distribution can be visualized to determine the degree of overlap or similarity. Bubble size is a measure of the topic prevalence. From the visualisation, we observe that (a) topics 1-3 and 3-2 overlap very slightly, while topic 4 and 5 are uniquely distinct. For the former, the overlap is statistically insignificant and thus, the 5 topics can be considered distinct enough to take this work forward. Furthermore, topic 1 appears to be the most prevalent topic, constituting 29.9% of the tokens. Figure 4 (right) is a sequence of stack bar graphs with the most prevalent words per topic. In the example shown in the image, the top 3 words to appear under topic 1 are “policy”, “technology”, “industry”, “government” and “environmental”, and are highlighted by the blue bars. However, in topic modelling, the relevance of a word is very important, as it leads to the topic being more obvious. This is where a relevancy metric, , is useful [15].

where is the probability of word w in topic k and is the lift in term’s probability within a topic to its marginal probability across the entire corpus. A lower λ gives more importance to the term , which gives more importance to topic exclusivity. Given in Figure 5 is the effect of reducing by 48%; the word “policy” remains unchanged, but interestingly, words like “industry” and “technology” swap places. Additionally, gone are the words “government” and “environmental” from the top 5 words within the topic, replaced instead with words like “firm” and “adoption”. (QN: NOTE TO SELF - WHY IS THIS IMPORTANT/ INTERESTING?. ALSO, is LAMBDA USED ANYWHERE ELSE IN THIS WORK??)

Chart, line chart

Description automatically generated

Figure 3: Coherence score as a function of topic number; generally, past 5 topics, the score does not vary significantly.

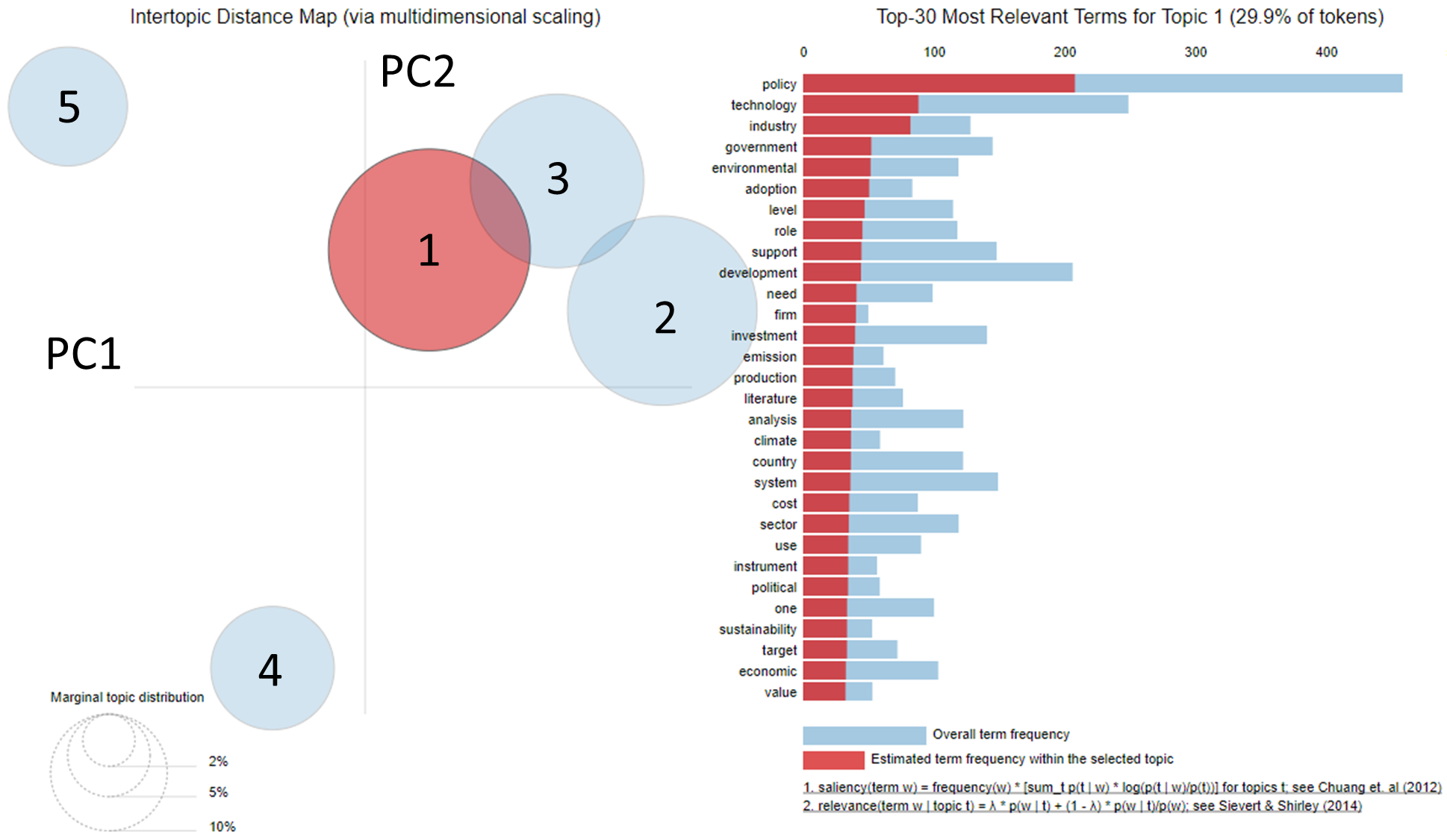


Figure 4: Topic Visualization using a Bubble map and the corresponding top 30 terms in each topic

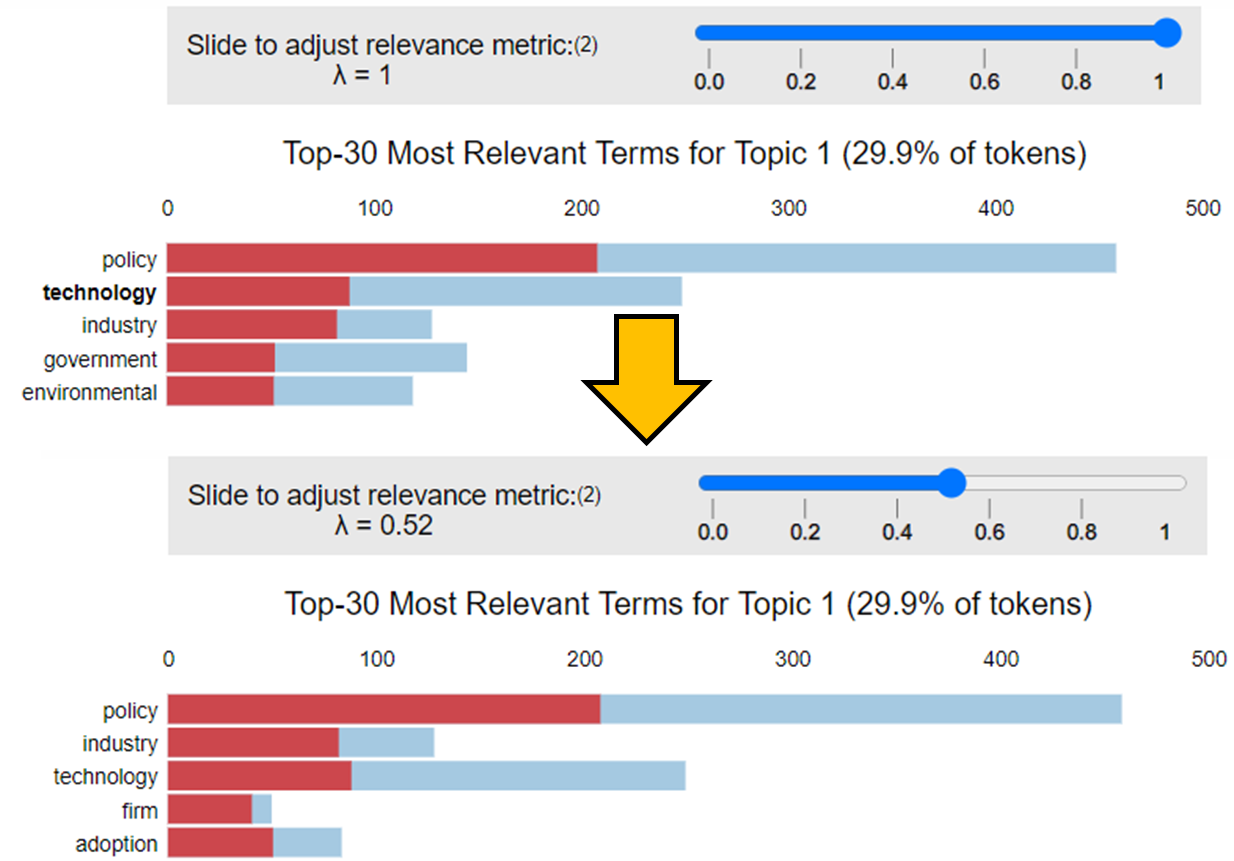


Figure 5: Topic Visualization using a Bubble map and the corresponding top 30 terms in each topic

**Topic Modelling with NNMF:** As another machine learning tool, NNMF is useful as it is able to dimensionally reduce the text data and create topics by grouping words with higher coherence [13]. It does this through a vector multiplication of “*W*” columns and “*H*” rows, with “*W*” representing the weightage of each word in a sentence (‘words-topics’ matrix) and “*H*” represents the words in each column (‘topics-documents’ matrix). The product, “*V*” is referred to as the ‘words-documents’ matrix. In factorizing “*V*” for uses in topic modeling, we are breaking down the corpus into “*W*” and “*H*”. As this approach is based on linear algebraic optimization, the outcome would be an approach that is independent of the probabilistic method utilized by LDA.

**Results:** Given in Table 1is the top 10 keywords per topic rank, sorted in order of importance. Each of the algorithms presents a slightly different result, although common repetitive words do occur within each algorithm (not surprising given the common data set). It is interesting to note the order of the words themselves and how each algorithm places importance on different words.

As an example, the results of the Gensim LDA within Rank 1 has sorted words: 'policy', 'investment', 'community', 'financial', 'performance', 'investor', 'system', 'institutional', 'technology' and 'measure'. The dark blue bars signify the weight of the word in the topic, while the lighter words signify the frequency of words. In topic 0 in figure 8, although the word “technology” has a much greater frequency than most words before it. It showed a lower weight for this specific topic. And words like “investment” is more representative of the topic due to its exclusiveness to the topic.

At face value, these words can be categorised as being more ‘quantitative’ in nature. With words like 'financial', 'performance' and ‘investor’, a subtopic of “financial performance” seems suitable.

This iterative process of (a) considering each line of output, (b) determining key words that help categorise the output as either ‘quantitative’ or ‘qualitative’, and (c) deciding on a subtopic that represents these words is perhaps the most subjective step of this entire process. The human element of interpreting, understanding and applying a judgement as to the most suitable topic, based on ones understanding of the words, is something that is unavoidable. Good topic models therefore require (apriori) that the modeler have some background knowledge of the particular subject matter.

A final step is to now consider all subtopics collectively and then generalize them into several key concepts, which than becomes the final topic model to be discussed in later sections. Generalisation removes some of the bias that may come about from subtopic selection, and allows for comparisons between the outputs of the algorithms.

Table 1: **Top 10 significand words per algoritm, proposed categorization, subtopic and key concept selection**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Rank of  Topic | Top 10 Words Weighted and Ranked | Categorisation | SubTopic | Key Concept |
| Gensim LDA | 1 | 'policy', 'investment' , 'community', **'financial', 'performance',**  'investor', 'system', 'institutional', 'technology', 'measure' | Quantitative | Financial Performance | Fiscal (Policy/Terms) |
| 2 | 'policy', 'technology', 'industry', **'government'**, 'environmental',  'adoption', 'level', 'role', 'support', 'development' | Qualitative | Environmental Technologies Policies | Government (Policy) |
| 3 | 'policy', **'stakeholder' , 'community'**, 'development', 'process', 'local', 'management', 'support', 'government', 'technology' | Qualitative | Community Development Policies | Community (Support/ Involvement) |
| 4 | '**technology'**, 'policy', 'system', 'development', **'community'**, **'traceability'**, 'group', **'social'**, 'support', 'case' | Quantitative-'Qualitative | Technology and social policies | Technology (Development) & Social (Policy) |
| 5 | **'state'**, 'policy', 'technology', **'investor'**, 'expansion', **'government'**, 'plant', 'country', 'electricity', **'community'** | Quantitative | Technological Policies and investors | Public-Private (Corisking) |
| Mallet LDA | 1 | sector', **'policy'**, 'electricity', 'development', 'country', 'source', **'production'**, 'support', **'data'**, 'emission' | Qualitative | Energy policies | Government (Policy) |
| 2 | technology', **'financial'**, 'impact', **'performance', 'cost', 'measure',** 'plant', 'adoption', 'system', 'importance' | Quantitative | Financial performance | Fiscal (Policy/Terms) |
| 3 | **policy', 'technology'**, 'investment', 'industry', **'environmental'**, 'investor', 'role', 'potential', **'framework'**, 'change' | Quantitative-'Qualitative | Technology investment policies | Technology (Development) & Social (Policy) |
| 4 | **stakeholder'**, 'process', **'social'**, 'management', **'community'**, 'model', **'relationship'**, 'literature', **'analysis'**, 'sustainable' | Quantitative | Stakeholder Perception | Public-Private (Corisking) |
| 5 | **community'**, 'local', 'system', 'government', **'development'**, 'support', **'transition'**, 'policy', **'national'**, 'state' | Qualitative | Community Involvement | Community (Support/ Involvement) |
| Non-Negative Matrix Factorization | 1 | **environmental'**, **'mix'**, 'state', **'government'**, 'country', 'development', 'technology', 'industry', 'sector', **'policy'** | Qualitative | Country Environmental policies | Government (Policy) |
| 2 | **process**', 'development', 'technology', **'management'**, 'program', **'partnership'**, 'social', **'beneficiary'**, 'local', **'community'** | Quantitative | Technology development | Public-Private (Corisking) |
| 3 | **influence', 'green',** 'satisfaction', **'relationship'**, 'proposed', **'network',** **'twomode',** 'method', 'retrofit', 'stakeholder' | Qualitative | Stakeholder Relationship Satisfaction | Community (Support/ Involvement) |
| 4 | market', 'retail', 'portfolio', 'policy', 'performance', **'financial',** 'technology', 'preference', **'investor', 'investment'** | Quantitative | Investor preference | Fiscal (Policy/Terms) |
| 5 | **expansion', 'fit',** 'power', 'state', 'price', 'promotion', 'electricity', **'capacity', 'plant', 'system'** | Quantitative-'Qualitative | Energy systems | Technology (Development) & Social (Policy) |

**Discussion and Implications: In general, this should include a brief (1-2 paragraph) introduction, followed by a** statement of the specific scope of the study, followed by results and then interpretations. Please avoid statements of future work, claims of priority, and repetition of conclusions at the end.

**Conclusions:** In general, this should include a brief (1-2 paragraph) introduction, followed by a statement of the specific scope of the study, followed by results and then interpretations. Please avoid statements of future work, claims of priority, and repetition of conclusions at the end.

**Main Text:** In general, this should include a brief (1-2 paragraph) introduction, followed by a statement of the specific scope of the study, followed by results and then interpretations. Please avoid statements of future work, claims of priority, and repetition of conclusions at the end.

Subheadings (“Results”, “Discussion”, or more specific subheadings, but not a leading “Introduction”) may be included in Research Articles or Reviews and should be brief and set off by a paragraph break. Up to three levels of subheadings may be used if warranted (bold for level one, bold and italic for level two, and italic for level three). Reports should not have subheadings.

All figures and tables should be cited in order (as, for example, “Fig. 1” and “Table 1”), including those in the Supplementary Materials (which should be cited as, for example, “fig. S1” and “table S1”). You may include line or page breaks if you would like to place figures within the text near where they are referenced. Please do not place figures in text boxes.

References should be cited in parentheses with an italic number (*1*). Multiple reference citations are separated by commas (*2*, *3*) or if a series, en dashes (*4–6*). References are cited in order by where they first are called out, through the text, text boxes, figure and table captions, reference notes and acknowledgments, and then the supplementary materials.

Equations can be included. Use MathType (recommended) or use the legacy equation editor in Word (Chose Insert > Insert Object > Word Equation). We do not recommend using the native Word equation editor. This can in some cases produce less reliable MathML, the online markup language we use, which may result in display errors. If you enter equations in simple LaTeX, check that they will convert accurately (Word 2007 and higher can convert simple LaTeX equations).

# **References**

|  |  |
| --- | --- |
| [1] | S. Jai, "Indian renewable firms underperformed but credit quality intact: Moody's," [Online]. Available: https://www.business-standard.com/article/companies/indian-renewable-firms-underperformed-but-credit-quality-intact-moody-s-121030901082\_1.html#:~:text=%22About%2015%2D20%20per%20cent,the%20underperformance%20respectively%2C%E2%80%9D%20says%20Abhishek. |
| [2] | T. Kealy, "Does an embedded wind turbine reduce a company’s electricity bill? Case study of a 300 kW wind turbine in Ireland.," *Journal of Business Ethics,* vol. 145, no. 2, pp. 417-428, 2015. |
| [3] | . S. Jacobsson and A. Johnson, "The diffusion of renewable energy technology: An analytical framework and key issues for research," *Energy Policy,* vol. 28, no. 9, pp. 625-640, 2000. |
| [4] | A. Jagadeesh, "Wind energy development in Tamil Nadu and Andhra Pradesh, India institutional dynamics and barriers — A case study.," *Energy Policy,* vol. 28, no. 3, pp. 157-168, 2000. |
| [5] | G. Shrimali, S. Goel, S. Srinivasan and D. Neslon, "Solving India’s renewable energy financing challenge: Which federal policies can be most effective?," CPI-ISB Series, 2014. |
| [6] | G. A. Bradshaw and J. G. Borchers, "Uncertainty as information: Narrowing the science-policy gap.," *Conservation Ecology,* vol. 4, no. 1, 2000. |
| [7] | J. West, I. Bailey and M. Winter, "Renewable energy policy and public perceptions of renewable energy: A cultural theory approach," *Energy Policy,* vol. 38, no. 10, pp. 5739-5748, 2010. |
| [8] | A. Masini and E. Menichetti, "Investment Decisions in the Renewable Energy Sector: An Analysis of Non-Financial Drivers," *Technological Forecasting and Social Change,* vol. 80, no. 3, pp. 510-524, 2013. |
| [9] | R. Maqbool, Y. Rashid, S. Sultana and Y. Sudong, "Identifying the critical success factors and their relevant aspects for renewable energy projects; An empirical perspective.," *Journal of Civil Engineering and Management,* vol. 24, no. 3, pp. 223-237, 2018. |
| [10] | D. Ramage, E. Rosen, J. Chuang, C. D. Manning and D. A. McFarland, "Topic modeling for the social sciences," *NIPS 2009 workshop on applications for topic models: text and beyond,* vol. 5, pp. 1-4, 2009. |
| [11] | M. W. Bickel, "Reflecting trends in the academic landscape of sustainable energy using probabilistic topic modeling," *Energy, Sustainability and Society,* vol. 9, no. 49, 2019. |

**Acknowledgments:** XXXXX

**Fig. 1.** **The figure caption should begin with a short descriptive statement of the entire figure followed by additional text.** Captions should be immediately after each figure. The primary callout of each figure part is indicated with a bold capital letter enclosed in parentheses [e.g., (**A**)]. Additional callouts are indicated the same way, but without the bold format. If you prefer, you can place both figures and captions logically throughout the text near where they are cited rather than at the end of the file (but not both). If a paragraph in the main text begins with the name of a figure, write out “Figure” in full (e.g., <para>“Figure 1 shows….”)

**Fig. 2.** **You can place graphics in-line above each caption.** Please do not use text boxes to arrange figures. High-resolution (preferably editable PDF or Adobe Illustrator format) figure files will be requested following review.

**Table 1.** **Start this caption with a short description of your table.** Format tables using the Word Table commands and structures. Do not create tables using spaces or tab characters.