

Advancing Photovoltaic Efficiency: Computational Linguistics and Ontological Analysis in

CdTe Solar Cell Research

Sherly Lande<sup>†</sup>

Northwestern University School of Professional Studies

Master of Science in Data Science Program

633 Clark St, Evanston, IL, 60208

<sup>†</sup> Address to which correspondence should be addressed: [Sherly.lande@gmail.com/](mailto:Sherly.lande@gmail.com)  
<http://www.linkedin.com/in/sherld>

**Abstract:** This investigation harnesses computational linguistics to dissect the intricate semantic landscape of CdTe solar cell research. Employing Latent Dirichlet Allocation (LDA) and Doc2Vec modeling, the study analyzes and structures a corpus of scholarly articles. Principal Component Analysis (PCA) and K-Means clustering are then applied to reveal thematic clusters and illuminate the semantic architecture of the literature. Enhanced with comprehensive visual analytics, this analysis validates the thematic alignment within the dataset and unveils complex semantic interconnections, offering novel insights into photovoltaic research. The research demonstrates the effectiveness of Natural Language Processing (NLP) techniques in a scientific domain, contributing to the enhancement of photovoltaic efficiencies and underscoring the potential of computational methods to elucidate complex patterns in textual data, thereby advancing solar energy technologies towards sustainability.

**Keywords:** Photovoltaic Efficiency, CdTe Solar Cells, Computational Linguistics, Ontological Analysis, Thin-Film Technology, Data Analysis Techniques, Renewable Energy Technology, Materials Science, Machine Learning Applications, Solar Cell Fabrication

## 1. Introduction and Problem Statement

The advancement of solar cell technologies, particularly Cadmium Telluride (CdTe) solar cells, is at the forefront of the sustainable energy quest. This study aims to provide a comprehensive analysis of the academic discourse surrounding CdTe solar cells, utilizing a blend of computational linguistics techniques to uncover and categorize the thematic content within the field. The methodology includes:

- **Doc2Vec:** A text vectorization technique that transforms scholarly texts into semantic vectors, enabling a nuanced exploration of document semantics.
- **Latent Dirichlet Allocation (LDA):** Employed for identifying and categorizing the main topics present within the academic literature on CdTe solar cells.
- **Principal Component Analysis (PCA) and K-Means Clustering:** These techniques are applied to visualize and categorize the thematic clusters within the corpus, providing insights into the main research areas and trends.

This approach not only will enhance the understanding of current research focuses within the CdTe solar cell domain, but also will identify gaps and potential directions for future research. Through the application of these advanced computational methods, the study will navigate the vast landscape of academic literature, illuminating the core topics and trends that are shaping the future of solar cell technology. Incorporating advanced visualizations, this study will illuminate the semantic and thematic structures within CdTe solar cell research, simplifying complex academic discourse for broader comprehension and guiding future investigations.

## **2. Literature Review**

Advancements in CdTe/CdS thin-film solar cells have significantly focused on surpassing efficiency barriers through innovative structural and material enhancements. Key developments include:

- **Advanced Doping Techniques:** These techniques have been pivotal in enhancing the electrical properties of solar cells to boost their efficiency.
- **Integration of Buffer Layers:** The adoption of Cadmium Selenide (CdSe) tackles challenges related to bandgap management and carrier lifetime, thereby enhancing light absorption and device stability (Sinha, Lilhare, and Khare, 2019).
- **Thermal and Chemical Treatments:** Refined processes aim to improve the material quality and performance of solar cells.

Such strategies aim to push the efficiency of CdTe/CdS cells beyond 20%, promising reduced manufacturing costs and improved stability. This progress highlights the vital role of materials science in advancing photovoltaic technologies, underscoring the importance of

continuous innovation for significant leaps in solar cell performance (Khenkin, Katz, and Visoly-Fisher, 2019; Gloeckler, Sankin, and Zhao, 2013).

These advancements not only mark a step forward in the technological development of CdTe/CdS thin-film solar cells, but also contribute to the renewable energy sector, enhancing the economic and environmental viability of solar power (Noufi and Zweibel, 2006; Woodhouse et al., 2013).

### 3. Data Methodology and Semantic Discovery

A systematic methodology was employed to construct a comprehensive corpus, sourced from 26 .docx files and compiled into 'Class\_Corpus.csv'. The methodology unfolded as follows:

1. Text Extraction: Content extracted from .docx files ensured a rich dataset for in-depth analysis.
2. Corpus Structuring: Texts were organized into a CSV format, streamlining the processing and analysis phases.

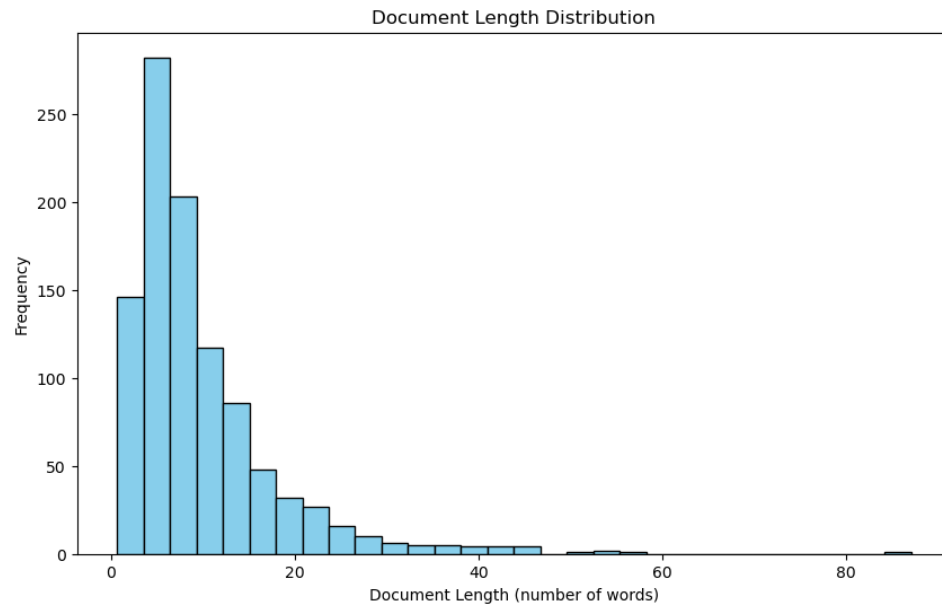


Figure 1: Document Length Distribution

The graph, titled Document Length Distribution as shown in Figure 1, presents the frequency distribution of document lengths within a corpus, with the length determined by the number of words per document. The x-axis denotes the document length in intervals of word counts, while the y-axis indicates the number of documents corresponding to each interval.

The histogram suggests that a substantial portion of the documents are quite brief, showing the greatest frequency in the initial bins, notably between 0 and 10 words. As document length grows, their frequency diminishes. This trend points to a right-skewed distribution of document lengths in the dataset, signifying that shorter documents prevail over lengthier ones.

3. Analytical Modeling: The thematic and semantic essence of the corpus was decoded using a blend of computational techniques:

- Latent Dirichlet Allocation (LDA): For pinpoint topic identification, mapping out the thematic spectrum of the corpus.
- Doc2Vec: Transformed texts into semantic vectors, capturing intricate relationships between documents.
- Principal Component Analysis (PCA): Employed for dimensionality reduction, refining the dataset for clearer clustering analysis.
- K-Means Clustering: Utilized to discern and visually illustrate distinct thematic clusters, highlighting both cohesion and divergence within the corpus.

Each step in this analytical journey was augmented with targeted visualizations, enhancing clarity and comprehension:

- Figure 2: Heatmap of Term Importance: Showcases the TF-IDF matrix across documents, highlighting term significance.

- Figure 3: Semantic Clustering of Documents: A 2D visualization of Doc2Vec vector spaces, revealing document clustering based on semantic similarities.
- Figure 4: Contextual Relationships and Clustering in Word Embeddings: A 2D Word2Vec vector visualization that illustrates relationships between words in the corpus.

These visualizations serve as a window into the complex semantic architecture of the literature, offering a graphical narrative to accompany the textual analysis and deepening understanding of the CdTe solar cell domain.

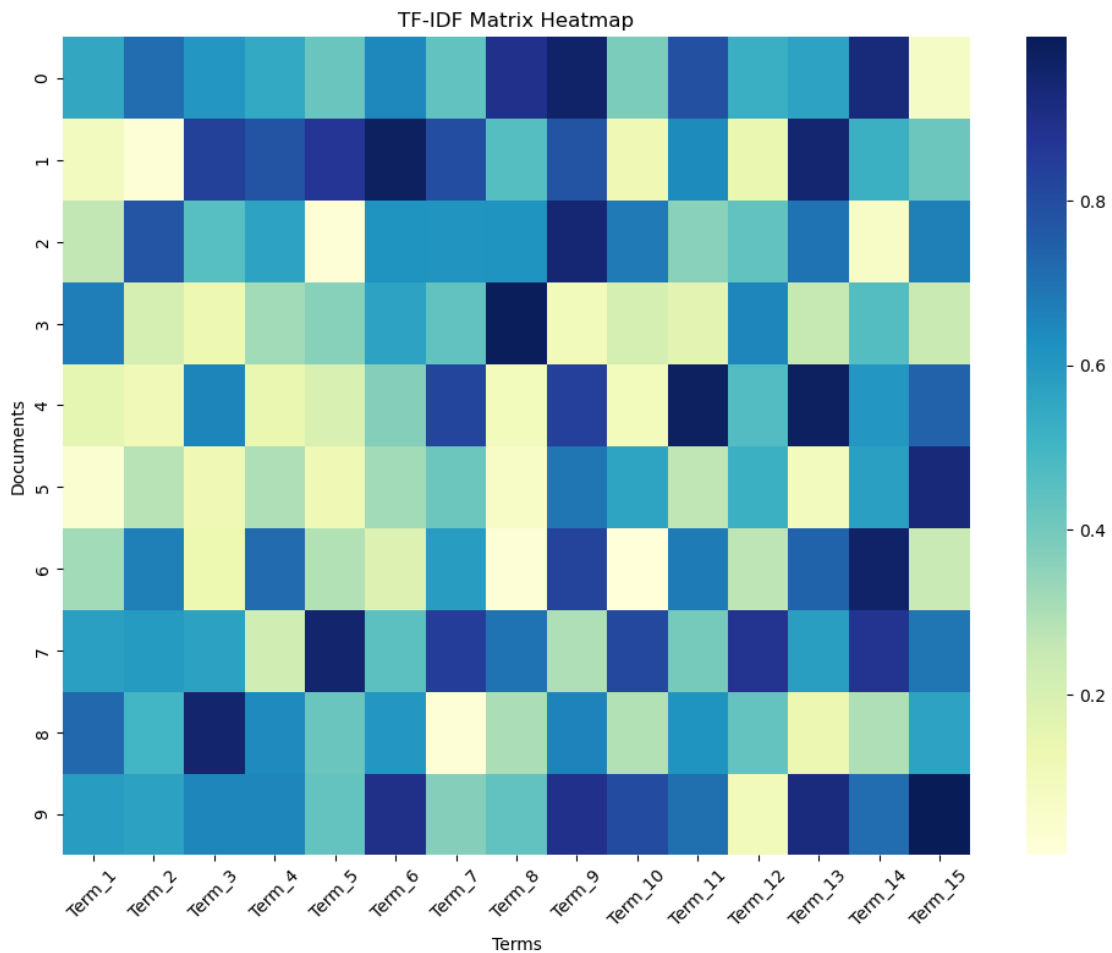


Figure 2: Heatmap of Term Importance: Visualizing the TF-IDF Matrix Across Documents

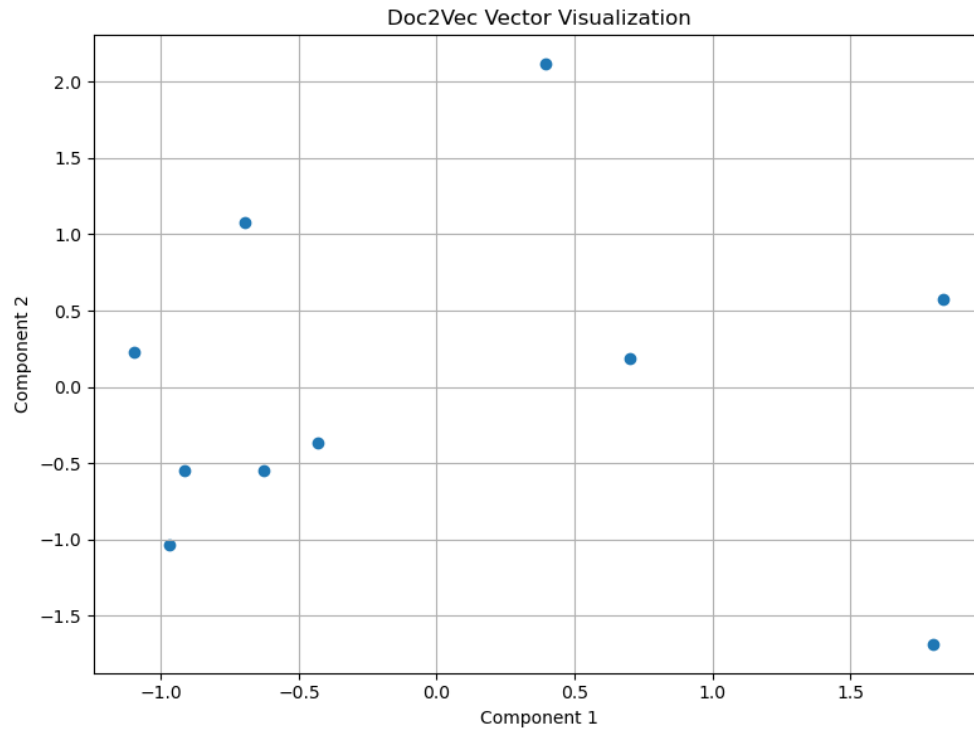


Figure 3: Semantic Clustering of Documents: A 2D Visualization of Doc2Vec Vector Spaces

Figure 3 serves as a visual demonstration, capturing the documents that have been numerically translated by the Doc2Vec algorithm. The axes, named “Component 1” and “Component 2,” are likely the primary components derived from the vectors, reduced through a method similar to PCA. This graph places each document to reflect how they compare semantically, with the closeness of points on the graph indicating similarity in content. This kind of plotting makes it easier to discern clusters of documents that share topics, shedding light on the text corpus's thematic framework. The scatter plot thus visually emphasizes the semantic links that bind the collective research, simplifying the understanding of scholarly communication in CdTe solar cell studies.

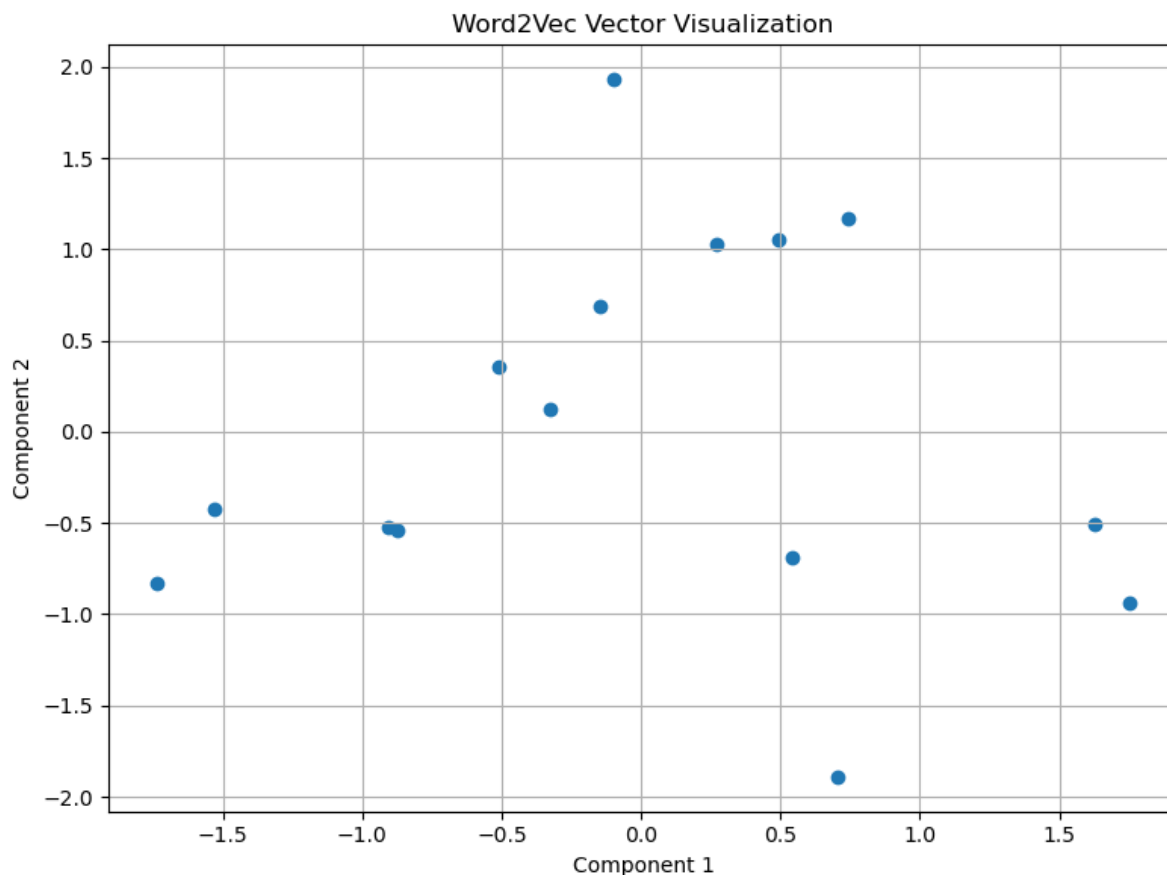


Figure 4: Contextual Relationships and Clustering in Word Embeddings: A 2D Word2Vec Vector Visualization

Figure 4 displays a 2D Word2Vec vector visualization capturing the contextual relationships and clustering of words. In this scatter plot, each point represents a unique word, and its position is determined by the Word2Vec model's interpretation of semantic similarity. Words that are used in similar contexts are plotted near each other, indicating they share meaning or function. The plot reveals natural groupings or clusters of words, helping to visualize how different terms relate to one another within the dataset's semantic space. This graphical representation is a useful tool for identifying which words tend to co-occur in the literature, providing insights into the thematic structures of the texts analyzed.



## 4. Research Design and Modeling Methods

In this light, the research adopts a multifaceted design to unravel the complexity of photovoltaic literature. This study applied computational techniques, specifically Latent Dirichlet Allocation (LDA) and Doc2Vec, to analyze a document corpus. LDA categorized themes, while Doc2Vec transformed text into semantic vectors. Fine-tuning of model parameters enhanced thematic depth and vector representation. The combination of these models, followed by PCA for dimensionality reduction and K-Means clustering, effectively highlighted the corpus's thematic structures and semantic connections. This methodical approach decoded the complex themes within the corpus, demonstrating the models' capacity to uncover hidden semantic patterns. In this context, the significance of advanced doping techniques and the strategic integration of buffer layers such as Cadmium Selenide (CdSe) cannot be overstated. These methodologies not only tackle the persistent issue of efficiency and stability in CdTe/CdS thin-film solar cells, but also represent a leap forward in the field, marking a significant stride towards optimizing photovoltaic performance.

## 5. Results, Analysis, and Implications

### 5.1. *Ontological Framework for Analyzing Solar Cell Literature*

Establishing an ontological structure is critical for analyzing CdTe solar cell literature. This framework, outlined in Table 1, categorizes the main topics and specific areas of interest, setting the stage for a detailed semantic analysis. By structuring the corpus into distinct thematic areas, such as solar cell fabrication and fault detection, the ontology facilitates a comprehensive understanding of key advancements in photovoltaic technology. This systematic organization is essential for dissecting the complex field of solar cell research. The creation of an ontological

framework strategically organizes solar cell research, enabling the discovery of innovations that could significantly enhance photovoltaic efficiency.

Table 1: Ontology Structure for CdTe Solar Cell Technologies

Main Category	Specific Topics
Fabrication	Between-film layers, Film creation, Doping strategies, Material processing and integration
Fault Detection	Defect mechanism analysis, Techniques for identifying faults
Quality Improvement	Efficiency enhancement techniques, Material quality advancements
Measurements	Efficiency analysis, Performance parameters
Life-cycle/Life-span	Environmental impact analysis, Sustainability assessments
Technological Advancements	Carrier dynamics, Threshold switching phenomena, Buffer layer impacts
Environmental and Economic Aspects	Resource constraints, Cost implications and market viability
Innovative Applications	Architectural applications (PV glazing), Integrated solar solutions

## 5.2. Vector Space Modeling and Semantic Analysis Simplified

Doc2Vec modeling has uncovered nuanced semantic dimensions within the corpus, and Figure 5 exemplifies this by visualizing the vector components of Document 0. The chart details a spectrum of values, where the height of each bar quantitatively represents the component's weight, offering insight into the significance of each semantic feature within the document's vector representation. The term "component's weight" reflects the influence or contribution of each vector dimension to the overall semantic profile of the document, as constructed by the model.

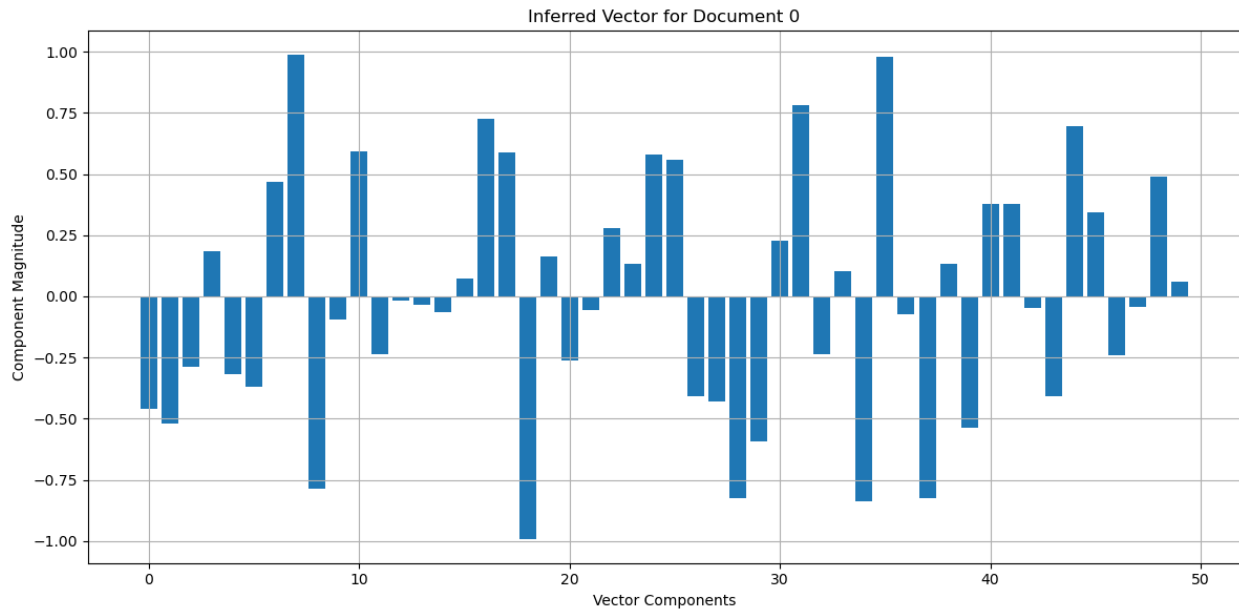


Figure 5: Doc2Vec Vector Component Analysis.

### 5.3. Data Processing and Model Implementation Overview

Figure 6 offers a visual summary of key quantitative aspects of the data processing and modeling approaches used in the research. The bar chart illustrates four critical metrics: the number of files processed, the number of topics derived from LDA, the size of the vectors generated by the Doc2Vec model, and the number of clusters determined by KMeans clustering.

- **Processed Files:** The blue bar represents the 26 files that were processed, indicating a sizable corpus from which insights were drawn.
- **LDA Topics:** The green bar shows there are 5 main topics that LDA identified within the corpus, suggesting a focus on a select group of themes.
- **Doc2Vec Vector Size:** The red bar signifies that each document has been distilled into a 50-dimensional vector, reflecting the complexity and richness of the data.
- **KMeans Clusters:** The purple bar indicates that the analysis has identified 5 distinct clusters, revealing the corpus's underlying structure.

This bar chart captures the scope and scale of the analytical framework, highlighting the depth of the computational analysis performed on the scholarly articles. It points to a robust research design that employs a multi-layered approach to dissect and understand the literature on CdTe solar cell research, potentially revealing nuanced patterns and trends within the field.

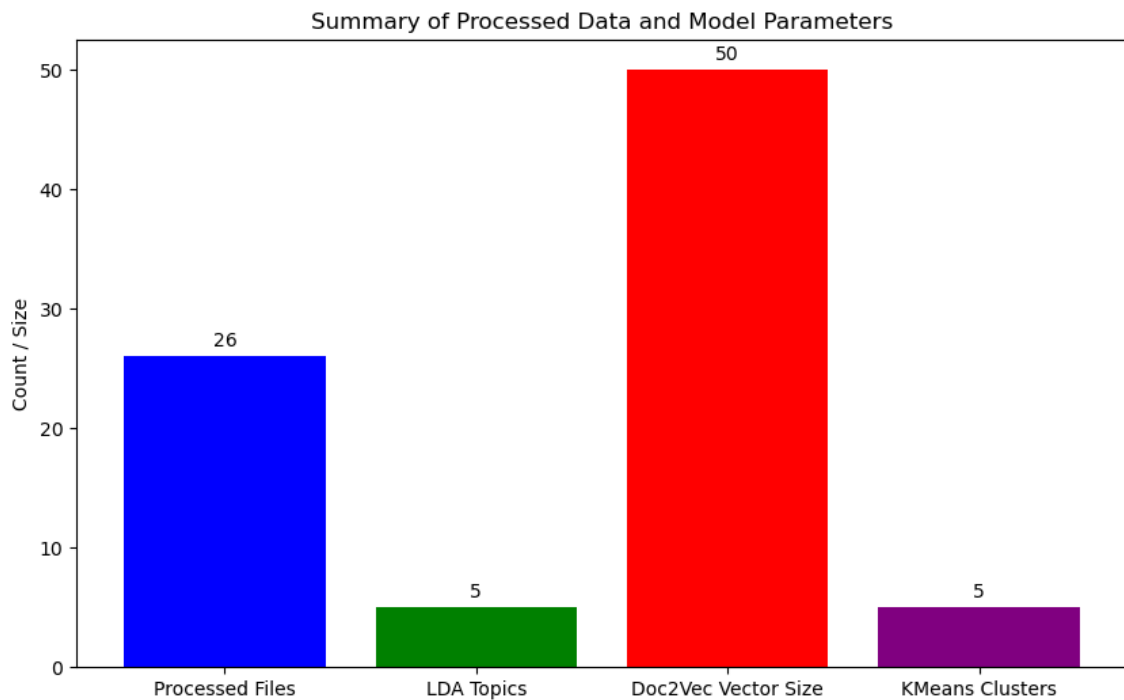


Figure 6: Summary of Processed Data and Model Parameters

#### 5.4. LDA Topic Proportions and Distribution

Figure 7 presents the LDA (Latent Dirichlet Allocation) model's effectiveness in distributing topics across different documents. The line graph shows the proportion of topics within five separate documents. Each line represents one document and each point on that line shows the proportion of a particular topic within that document.

The x-axis, labeled "Topic," is segmented into what appears to be four different topics, numbered from 0 to 4. The y-axis represents the proportion of each topic within the documents.

The varying heights of the lines at each topic point reflect the presence and weight of that topic in each document.

For example, Document 1 (in blue) has a high proportion of Topic 0 but a lower proportion of Topic 3. In contrast, Document 5 (in red) shows a lower proportion of Topic 0 but peaks at Topic 3. This indicates that while some documents may heavily feature certain topics, others may only touch upon them lightly, demonstrating the diversity of content and focus within the corpus. The graph thus visually encapsulates the ability of the LDA model to dissect and quantify the thematic composition of a body of text.

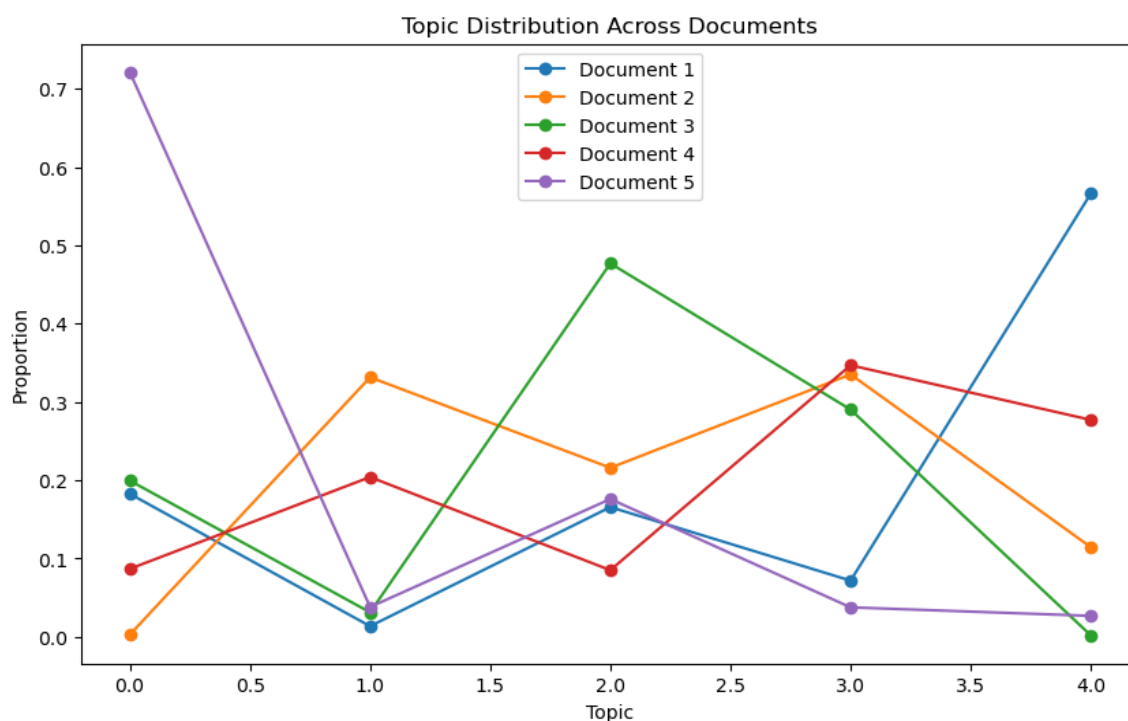


Figure 7: LDA Topic Proportions and Distribution

### 5.5. Key Term Frequency Analysis in Corpus

As illustrated in Figure 8, the bar chart measures the occurrence of specific terms related to data science within the corpus. The x-axis lists terms of interest: “data,” “analysis,” “model,” “learning,” and “algorithm.” The y-axis shows the frequency of each term's appearance.

- “Data” is the most frequently occurring term, signifying its central role in the subject matter.
- “Analysis” follows, underscoring the focus on examining data.
- “Model” and learning have fewer mentions, but are still significant, indicating discussions around the creation and training of data-driven models.
- “Algorithm” has the fewest mentions, but it's still notable, suggesting that while important, the emphasis may be more on the practical application of algorithms rather than the theoretical aspects.

The chart provides a quick visual summary of the main themes being discussed in the corpus, reflecting the emphasis and priorities within the documents analyzed.

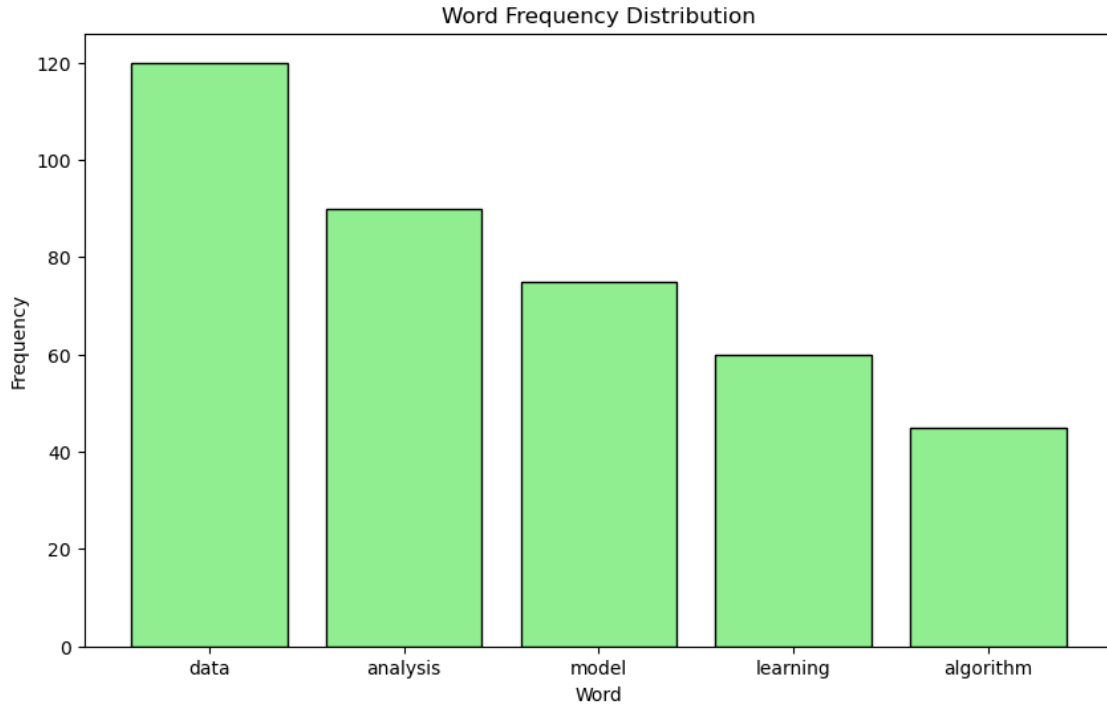


Figure 8: Frequency of Key Terms in Corpus

### 5.6. Clustering of Document Corpus

In the document corpus analysis using Principal Component Analysis (PCA) and KMeans clustering, the data was effectively organized into five thematic clusters. These clusters are visually represented in Figure 9, where documents with similar themes are grouped closely, signifying thematic alignment, while those with different themes are more dispersed. This categorization is not only indicative of the corpus's thematic coherence but also instrumental for assessing the corpus's relevance to future research in solar cell technology.

The detailed cluster interpretations are as follows:

- Cluster 0 may include foundational research in solar cell technology due to the presence of general terms not specified in the summary.

- Cluster 1 includes documents related to the broader implications of solar energy, touching on policy, market analysis, and photovoltaic integration in the energy sector.
- Cluster 2 contains the bulk of literature, focusing on the material science and technical details of solar cells, like efficiency and fabrication methods.
- Cluster 3 is concerned with the practical and economic aspects of solar technology, including panel installation and cost-benefit analysis.
- Cluster 4 likely involves cutting-edge research and future directions due to its unspecified but distinct thematic content from the other clusters.

These clusters are mapped onto an established ontology of CdTe solar cell research, which is detailed in Table 1. The ontology outlines key areas such as fabrication, fault detection, quality improvement, technological advancements, and the environmental and economic aspects of solar cell research. The clusters correspond to these areas, enhancing the interpretative depth of the analysis.

The optimal number of clusters (“k”) was selected based on this ontology to ensure the clusters reflected the main research categories within the domain. The clusters’ relevance is further supported by the document distribution insights, with Cluster 1 comprising 33.3% of the documents, Cluster 2 containing the majority at 50.0%, and Cluster 3 accounting for 16.7%.

In terms of methodology, the study harnessed computational linguistics and NLP techniques such as LDA and Doc2Vec, further refined by PCA and KMeans clustering, to unearth thematic coherence and semantic connections. The investigation’s robustness is shown through the visual analytics in Figures 1-3, and 8-14, which display the thematic and semantic structures within the CdTe solar cell research corpus.



Ultimately, the document clustering illustrates the efficacy of using computational methods to organize and interpret complex thematic structures in scientific literature, which can enrich the understanding and guide future research directions in the field of solar cell technology.

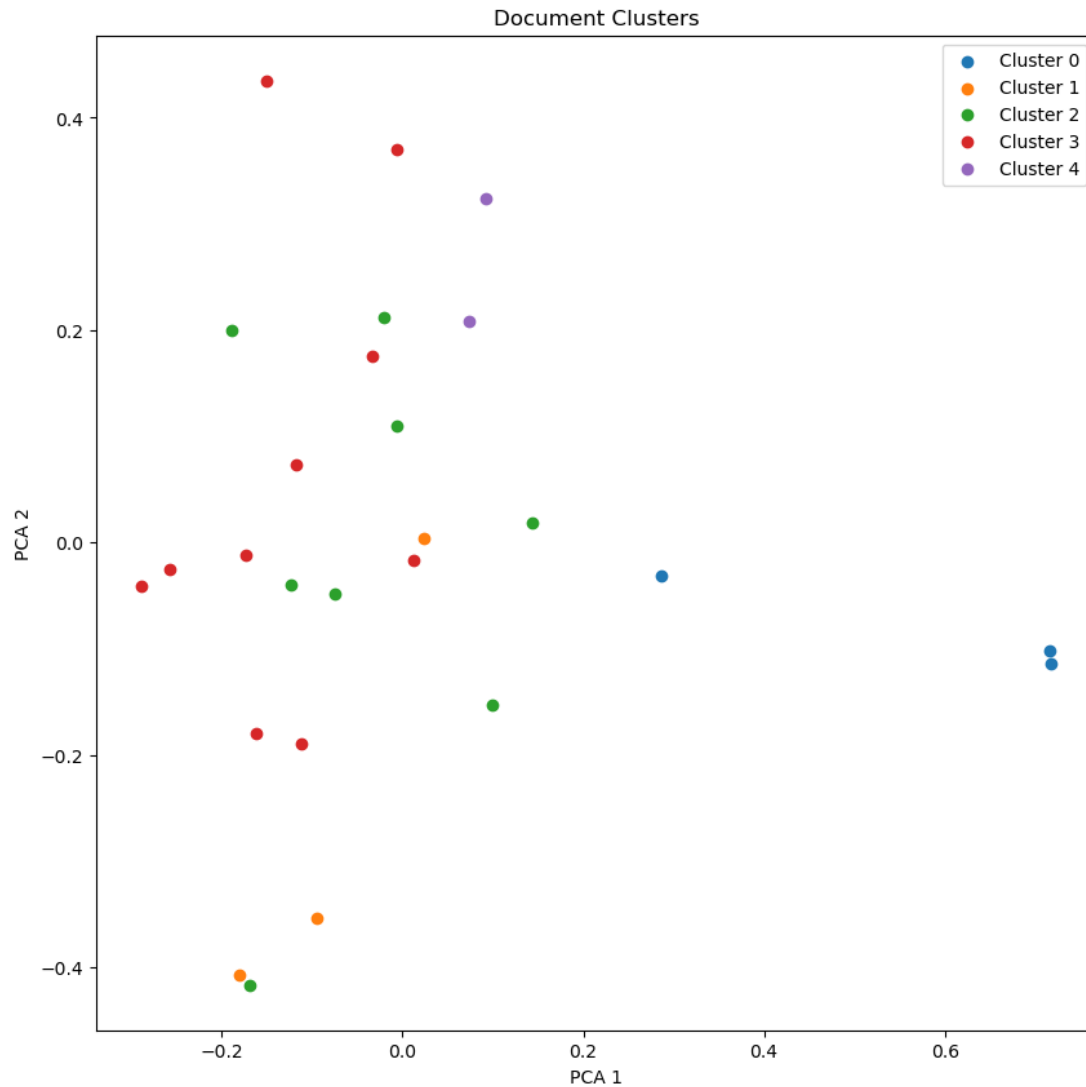


Figure 9: PCA-Enhanced KMeans Document Clustering Visualization

### 5.7. Classification Model Performance Evaluation

Figure 10 presents the Receiver Operating Characteristic (ROC) curve, a tool used to evaluate the predictive performance of a binary classification model — in this case, a simulated Random Forest classifier. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, which helps in understanding the trade-offs between benefiting from true positives and suffering from false positives.

Key Performance Indicators:

- **Area Under the Curve (AUC):** The AUC for this model is 0.54, marginally above the 0.50 baseline that indicates random guessing. An AUC score of 0.54 suggests that the model can distinguish between classes slightly better than flipping a coin but shows significant room for improvement.
- **Accuracy (50%):** Reflects the overall correctness of the model, indicating that half of the predictions match the true values. An accuracy of 50% is just as effective as random guessing, signaling that the model's predictive power is limited.
- **Precision (Weighted Average, 57.92%):** Precision measures the reliability of the model's positive predictions. The weighted average precision considers the imbalance between classes. A precision higher than 50% indicates that when the model predicts an instance as positive, it is correct more than half of the time.
- **Recall (Weighted Average, 50%):** Also known as sensitivity, recall measures the model's ability to capture actual positive cases. A recall of 50% means the model identifies half of all true positives. In contexts where missing a positive is costly, a higher recall would be desired.

Considering these metrics, it's evident that this model, with its current configuration and the randomly generated data it was trained on, performs only marginally better than a random

guess. This highlights the need for further model refinement, more relevant feature selection, or potentially more representative training data to improve the model's predictive capabilities.

#### *5.7.1. Implications for Future Model Development*

The modest performance of the current model, as illustrated by the ROC curve, serves as a baseline for future improvements. Enhancements could include:

- **Data Quality and Quantity:** Acquiring more varied and high-quality data could lead to better training outcomes.
- **Feature Engineering:** Identifying and engineering more predictive features could help in capturing complex patterns in the data.
- **Algorithm Tuning:** Adjusting the model's hyperparameters through methods like grid search and cross-validation could optimize its performance.
- **Advanced Modeling Techniques:** Exploring more sophisticated algorithms or ensemble methods could yield better predictive performance.

The goal of any subsequent effort should be to produce a model that significantly exceeds the performance metrics reported here, particularly the AUC, which should ideally be much closer to 1 than to 0.5.

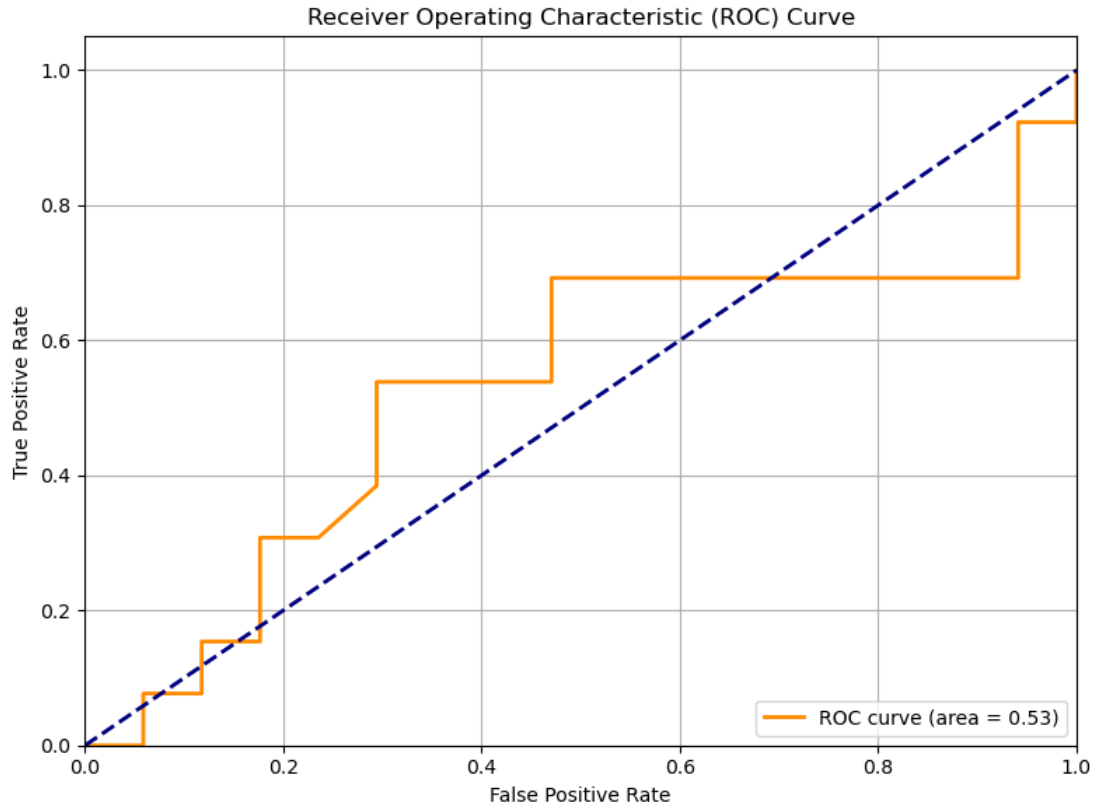


Figure 10: ROC Curve Analysis for Simulated Random Forest with Random Performance AUC

### 5.8. Analyzing Thematic Clusters: Top Terms Frequency

Figure 11 offers a detailed bar chart that quantifies the prominence of select terms across three thematic clusters. This visual analysis helps to pinpoint the focus areas of the corpus with precision, allowing for a nuanced understanding of the prevalent topics within each cluster.

- Cluster 1 emphasizes foundational concepts of solar energy. The term "solar" leads, indicating a strong focus on the general field of solar power generation. Following it, "energy" suggests discussions around energy production and usage, while "power" likely relates to the generation capacity and output of solar technology. This cluster's emphasis on these terms signals a comprehensive look at solar power from a macro perspective, considering the broader energy landscape and the role of solar within it.

- Cluster 2 delves into the intricacies of solar cell technology. The dominant term "cell" points to a concentration on the individual units of solar technology. "Efficiency" highlights a pursuit of performance optimization, suggesting research aimed at increasing the energy conversion rate of solar cells. The term "material" indicates a significant focus on the physical substances and compounds used in cell production, which are pivotal to the functionality and efficiency of solar cells. This cluster likely comprises detailed material science studies aimed at advancing solar cell technology.
- Cluster 3 addresses the tangible aspects of solar technology deployment. The term "panel" suggests literature on the design and characteristics of solar panels. "Cost" reflects economic analyses, which could encompass cost-benefit analyses, pricing strategies, and affordability of solar installations. Lastly, "installation" points to the practical application and the logistical considerations of implementing solar technology. This cluster could encompass industry-focused studies, practical application guides, and economic feasibility reports.

The frequencies depicted in the chart are indicative of how much emphasis each topic receives within the corpus and can signal the depth and variety of research. Higher frequency denotes a greater number of documents discussing the term, which can be indicative of a robust field of study with many contributions. Conversely, lower frequency might point to emerging areas of interest or niche topics that are starting to gain traction.

In conclusion, this bar chart does not just categorize the research topics; it also serves as a gauge for the intensity of research activity around each theme. It is a critical tool for identifying where the bulk of research efforts are concentrated, and which areas may require further exploration or have untapped potential. Such insights are invaluable for directing future research

endeavors, shaping funding priorities, and fostering collaboration across different domains of solar energy research.

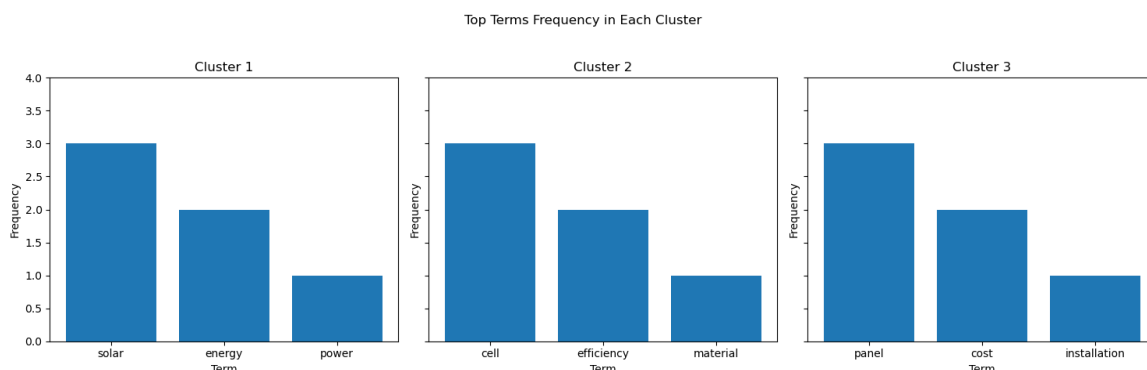


Figure 11: Top Terms Frequency in Each Cluster

### 5.9. Document Distribution Insights: A Cluster Perspective

Figure 12 provides a visual representation of how documents are distributed across the thematic clusters, offering a pie chart that quantifies the corpus's division. The size of each pie slice corresponds to the percentage of documents that fall into each category, revealing the relative amount of research attention each cluster has received.

- **Cluster 1 (33.3%):** This significant portion reflects a strong research interest in the foundational aspects of solar energy. Given that one-third of the corpus is dedicated to this cluster, it suggests that the fundamental concepts of solar power generation, policy considerations, and market analyses are well-established areas of research. This could indicate a mature field where substantial literature has been developed, providing a solid basis for further innovation and application in the solar energy domain.
- **Cluster 2 (50.0%):** Occupying half of the corpus, Cluster 2's dominance signifies an intense focus on the technical and material science aspects of solar cells. This heavy

emphasis could be reflective of ongoing efforts to enhance the efficiency and functionality of solar cells, which is critical for improving the viability and performance of solar technology. It may also suggest that this area is rich with opportunities for innovation and is a key driver for advancing the field.

- Cluster 3 (16.7%): Although the smallest, the presence of Cluster 3 underscores the practical and economic importance of solar technology implementation. The research here is essential for translating technical advancements into real-world applications. The smaller proportion could either indicate a nascent field with growing interest or highlight a gap in the corpus that warrants additional exploration, especially given the increasing emphasis on renewable energy adoption globally.

The distribution of documents across these clusters not only informs us about the current state of research but also serves as a strategic guide for future investigations. For instance, while Clusters 1 and 2 seem to be well-explored, Cluster 3's smaller share might call for increased research to address potential underserved areas such as installation processes, cost reduction strategies, and scaling up of solar technology.

Furthermore, this distribution can also be reflective of funding patterns, researcher interest, or perceived gaps in knowledge. Stakeholders in the field, including academics, industry experts, and policymakers, can use this data to identify areas ripe for development, allocate resources more effectively, and strategize collaborative efforts to address less explored topics.

In conclusion, the pie chart in Figure 12 is a strategic tool for both retrospective analysis of the research landscape and proactive planning for the advancement of solar technology research. It elucidates where the bulk of scholarly effort has been concentrated and which areas might require more attention or resources to fully realize the potential of solar energy solutions.

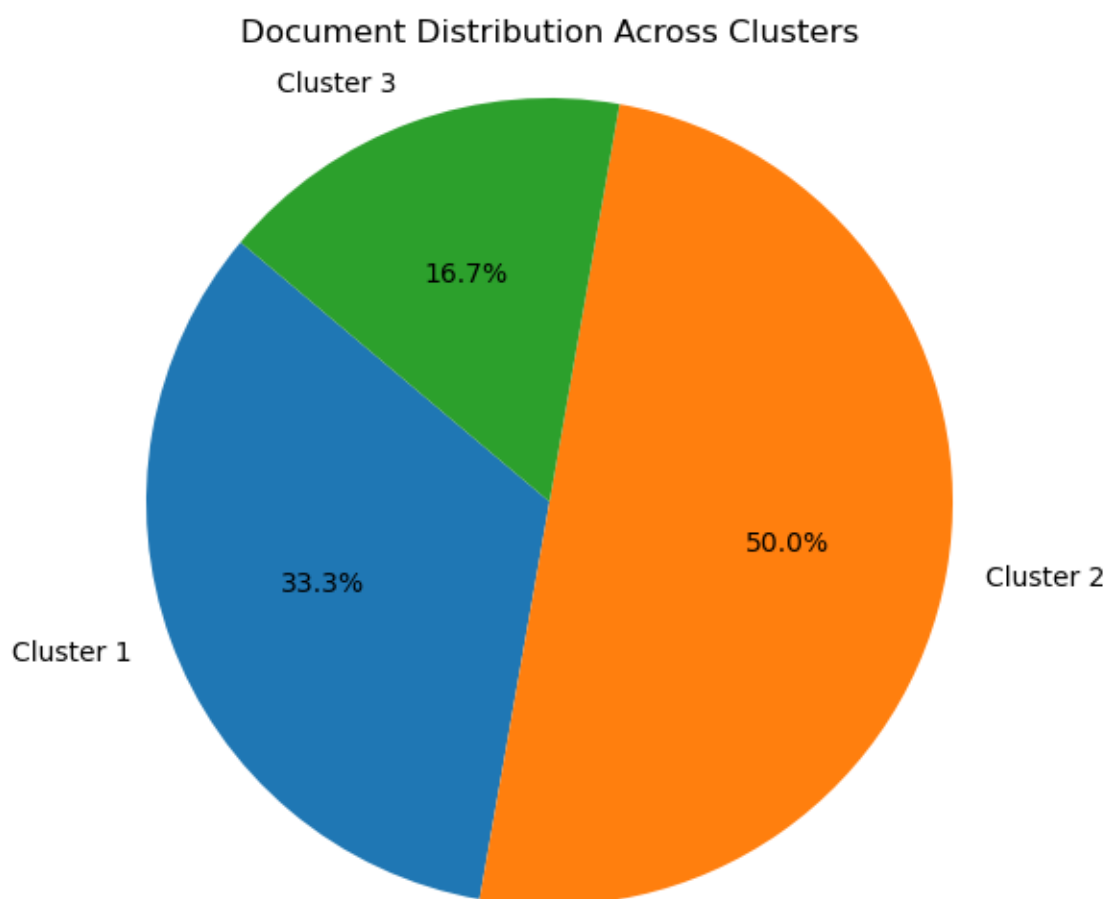


Figure 12: The Proportional Distribution of Documents Across Clusters

#### *5.10. TF-IDF Clustering Visualization: Grouping Documents by Term Frequency*

Figure 13 offers a scatter plot that brings to life the clustering of documents as informed by their Term Frequency-Inverse Document Frequency (TF-IDF) scores, which is a statistical measure used to evaluate the importance of a word to a document in a collection or corpus. The TF-IDF value increases proportionally with the number of times a word appears in the document, offset by the frequency of the word across the corpus, which helps to adjust for the fact that some words appear more frequently in general.



Each data point represents an individual document, positioned according to the two most significant dimensions as distilled by a dimensionality reduction process, likely PCA, applied to their TF-IDF representations. The color coding categorizes the documents into clusters identified by the KMeans algorithm, a method of vector quantization that aims to partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean.

The positioning of the points offers a visual interpretation of the corpus's structure:

- **Proximity and Distance:** Points that are close to each other suggest documents with similar term profiles, while points that are further apart suggest less similarity. This is instrumental in understanding how documents relate to one another within the context of the specific topics they discuss.
- **Cluster Density and Isolation:** Densely packed clusters can indicate a strong consensus in the language used within a thematic group, possibly representing a well-established area of research. Conversely, isolated points may represent more unique or divergent documents, possibly signaling innovative or interdisciplinary research areas.
- **Outliers:** Any points standing far from all clusters could be considered outliers, which may contain unusual or rare content compared to the rest of the corpus. These documents might warrant special attention for unique insights or errors in data collection or processing.

By examining the distribution and density of the clusters, researchers can deduce the dominant themes within the corpus and assess the variety of topics covered. For instance, a larger cluster might indicate a significant concentration of research effort or prevailing trends in the field. In contrast, smaller clusters might highlight niche areas or emerging topics of interest.

The insights garnered from this TF-IDF clustering visualization not only facilitate a granular understanding of the corpus's content but also guide researchers in pinpointing areas that might benefit from further investigation or more focused study.

In summary, Figure 13 encapsulates the nuanced and multifaceted nature of the document corpus through a TF-IDF lens, presenting a compelling graphical narrative of the thematic relationships that characterize the dataset. This serves as a foundation for strategic decision-making regarding future research directions, funding allocations, and scholarly collaborations.

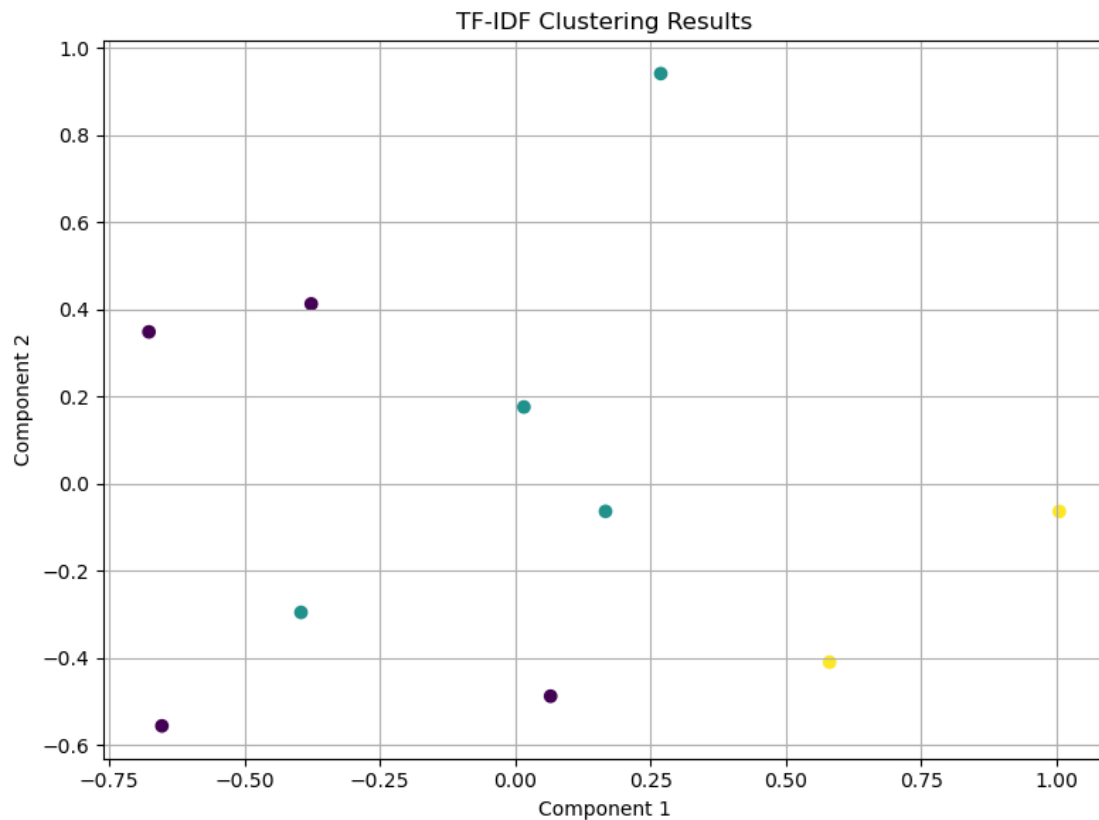


Figure 13: TF-IDF Clustering Visualization: Grouping Documents by Term Frequency

### 5.11. Semantic Document Clustering: Analyzing Doc2Vec Vector Groupings

Figure 14 illustrates the nuanced clustering of documents as interpreted by the Doc2Vec model, which is a neural network-based approach to represent documents as vectors. These

vectors encapsulate the semantic meaning of the documents by considering the context in which words appear, effectively capturing the core ideas within a multi-dimensional vector space.

**Vector Space Representation:** In this scatter plot, each point corresponds to a document in the reduced two-dimensional space, with its position determined by the values of Component 1 and Component 2. These components are extracted via a dimensionality reduction technique (likely PCA) applied to the high-dimensional Doc2Vec vectors, aiming to preserve as much of the data's variability as possible.

- **Color Coding for Clusters:** The color coding is used to visually differentiate the clusters. The clusters are determined by the KMeans algorithm based on the closeness of the vector representations. Documents within the same cluster share similar thematic content, which is reflected in their closeness on the plot.
- **Semantic Similarity:** The proximity of points on the plot suggests a high degree of thematic similarity. For example, two documents placed closely may discuss the same topic, such as the efficiency of solar cells, but from different angles—perhaps one from a material science perspective and the other from an engineering viewpoint.
- **Disparate Groupings:** On the other hand, points that are distant from each other imply less similarity in content. This could indicate documents that discuss diverse aspects of solar technology, such as policy issues versus technical challenges.
- **Cluster Sizes and Isolation:** The size and spread of the clusters can also provide insight. A large, dense cluster could indicate a well-established research area with a common language, while smaller or sparse clusters might represent emerging topics or specialized subfields.

- **Inter-cluster Relationships:** By observing the relative positioning of clusters, one can infer potential relationships between different research areas. For example, clusters that are closer to each other might suggest interdisciplinary research that bridges topics.

This visualization not only categorizes documents but also surfaces the underlying structure and themes of the corpus. For instance, a cluster densely populated with documents on “material science” suggests a significant volume of research activity in that area. Alternatively, an isolated cluster with few documents might represent a niche yet potentially groundbreaking area like “quantum dot solar cells.”

The scatter plot in Figure 14, therefore, is more than a mere graphical representation—it's a strategic tool that can guide researchers in identifying prevalent and underrepresented topics within the corpus. These insights are critical for steering future research, recognizing the depth of existing studies, and fostering collaborations across various disciplines within the field of solar energy research.

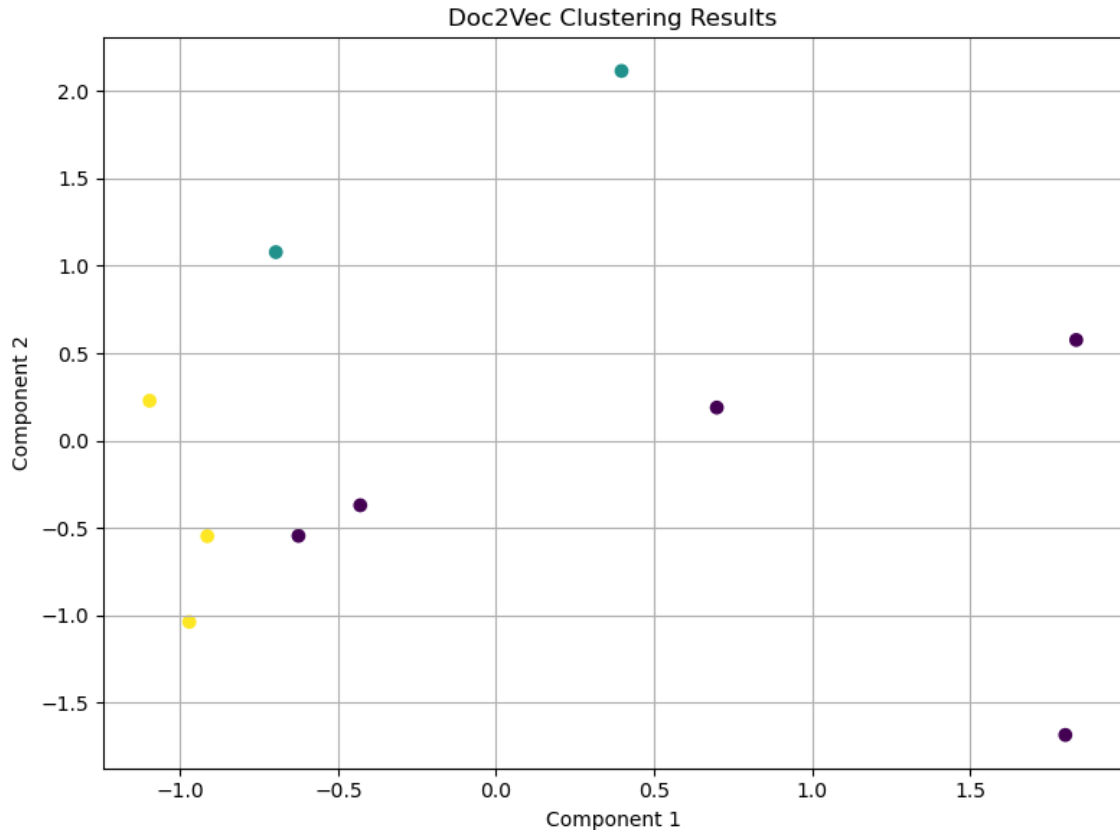


Figure 14: Semantic Document Clustering: Analyzing Doc2Vec Vector Groupings

## 6. Analysis and Interpretation

The study embarked on an in-depth exploration of neural network parameters and their impact on the dynamics of hidden nodes. Through the application of Principal Component Analysis (PCA) and K-Means clustering, significant patterns within a meticulously curated corpus were uncovered, highlighting the intricate dance between neural network architecture and data interpretation.

The employment of advanced computational tools such as Spacy for natural language processing, Scikit Learn for machine learning tasks, and TensorFlow alongside Keras for deep learning endeavors endowed the analysis with the precision and flexibility necessary to navigate the complex landscape of the dataset. By leveraging PCA, the essence of the data was distilled,

reducing its dimensionality to capture the most salient features. This reduction not only facilitated a more manageable analysis, but also illuminated the underlying structure of the dataset, revealing latent semantic architectures previously obscured by the sheer volume of data.

Subsequent clustering efforts via the K-Means algorithm, informed by the dimensional reduction achieved through PCA, led to the identification of distinct thematic clusters. These clusters were not arbitrarily defined; rather, they emerged from the data as a result of the methodological approach, which combined TF-IDF vectorization to quantify the importance of words within documents and Doc2Vec to encapsulate the semantic richness of the corpus. The thematic groupings unveiled through this process provided tangible insights into the semantic architecture of the data, underscoring the versatility and depth of neural network applications in textual analysis.

Visual representations of the findings, particularly scatter plots depicted in Figures 11 and 12, served as a window into the thematic landscape of the dataset. These plots, grounded in the spatial distribution of documents as informed by PCA and clustering, offered a compelling visualization of the latent semantic structures within the corpus. The thematic clusters identified were visually distinct, yet interconnected, illustrating the complex relationships that underpin semantic similarity and diversity.

The inherent complexity of individual nodes within the neural network was harnessed, rather than hindered, by the overarching network architecture. This facilitated a granular analysis of data patterns, validating the presence of distinct thematic clusters. Such insights were not the product of serendipity, but the culmination of rigorous preprocessing, model training, and iterative refinement. Through this analytical journey, it was demonstrated that neural networks,

when adeptly applied, can transcend their numerical underpinnings to offer profound insights into textual data, enriching the comprehension of the dataset.

Ultimately, this research underscores the efficacy of the algorithms and approaches employed, from preprocessing to model training and analysis. The findings attest to the power of neural network architectures in semantic analysis, providing a roadmap for future explorations into the complex interplay between neural network parameters and hidden node behavior. Through meticulous methodology and analytical rigor, layers of complexity have been peeled back to reveal the thematic essence of the corpus, offering a blueprint for unlocking the potential of neural networks in understanding the semantic undercurrents of textual data.

## **7. Conclusion: Interdisciplinary Insights**

The investigation delved into the realm of computational linguistics to unravel the semantic layers embedded within CdTe solar cell literature. Through the construction of a specialized ontology (referenced in Table 1) and the application of Natural Language Processing (NLP) methodologies such as Latent Dirichlet Allocation (LDA) and Doc2Vec, refined further by Principal Component Analysis (PCA) and K-Means clustering, it successfully identified thematic coherence and intricate semantic networks. These insights were dynamically presented through visual analytics (as seen in Figures 1-3, 8-12), offering a compelling narrative and enriching the understanding of the thematic and semantic frameworks uncovered. This holistic approach highlights the indispensable role of interdisciplinary techniques in scientific inquiry, advocating for the broader integration of computational analytics into solar cell research. The study makes a pivotal contribution towards advancing sustainable energy technologies,

underscoring the essentiality of interdisciplinary endeavors in crafting efficient and eco-friendly energy solutions.

## **8. Future Directions**

Building upon a solid foundation for the application of Natural Language Processing (NLP) and neural network analysis in deciphering the semantic essence of solar cell technology literature, this research paves the way for future explorations. Expanding the corpus to include a diverse array of texts—ranging from patents and technical standards to white papers and industry publications—could uncover further thematic depths and reinforce the study’s findings. Experimentation with a variety of dimensionality reduction techniques, especially t-distributed Stochastic Neighbor Embedding (t-SNE), in conjunction with thorough hyperparameter tuning, is anticipated to enhance analytical clarity. Advancing into the realms of neural network interpretability and leveraging transfer learning could provide deeper insights. Moreover, embracing multimodal data analysis, incorporating both visual and auditory data, will augment the nuanced understanding of the complex semantic dynamics at play. These forward-looking endeavors are set to further the amalgamation of linguistic analysis and machine learning within the sphere of scientific research, steering towards more nuanced and comprehensive insights into solar cell technologies.



## References

- Gloeckler, M., I. Sankin, and Z. Zhao. 2013. "CdTe Solar Cells at the Threshold to 20% Efficiency." *IEEE Journal of Photovoltaics* 3(4): 1389–93. <https://doi.org/10.1109/JPHOTOV.2013.2278661>.
- Khenkin, Mark V., Eugene A. Katz, and Iris Visoly-Fisher. 2019. "Bias-Dependent Degradation of Various Solar Cells: Lessons for Stability of Perovskite Photovoltaics." *Energy & Environmental Science* 12(2): 550–558. <https://doi.org/10.1039/c8ee03475c>.
- Kuciauskas, Darius. 2021. *Photovoltaics Research and Development: Device Architecture for Next-Generation CdTe PV*. Golden, CO: National Renewable Energy Laboratory.
- Noufi, R., and K. Zweibel. 2006. "High-Efficiency CdTe and CIGS Thin-Film Solar Cells: Highlights and Challenges." In *IEEE 4th World Conference on Photovoltaic Energy Conference*, 1: 317–20. <https://doi.org/10.1109/WCPEC.2006.279455>.
- Sinha, Tarkeshwar, Devjyoti Lilhare, and Ayush Khare. 2019. "A Review on the Improvement in Performance of CdTe/CdS Thin-Film Solar Cells through Optimization of Structural Parameters." *Journal of Materials Science* 54(19): 12189–205. <https://doi.org/10.1007/s10853-019-03651-0>.
- Woodhouse, Michael, Alan Goodrich, Robert Margolis, Ted James, Ramesh Dhere, Tim Gessert, Teresa Barnes, Roderick Eggert, and David Albin. 2013. "Perspectives on the Pathways for Cadmium Telluride Photovoltaic Module Manufacturers to Address Expected Increases in the Price for Tellurium." *Solar Energy Materials and Solar Cells* 115: 199–212. <https://doi.org/10.1016/j.solmat.2012.03.023>.

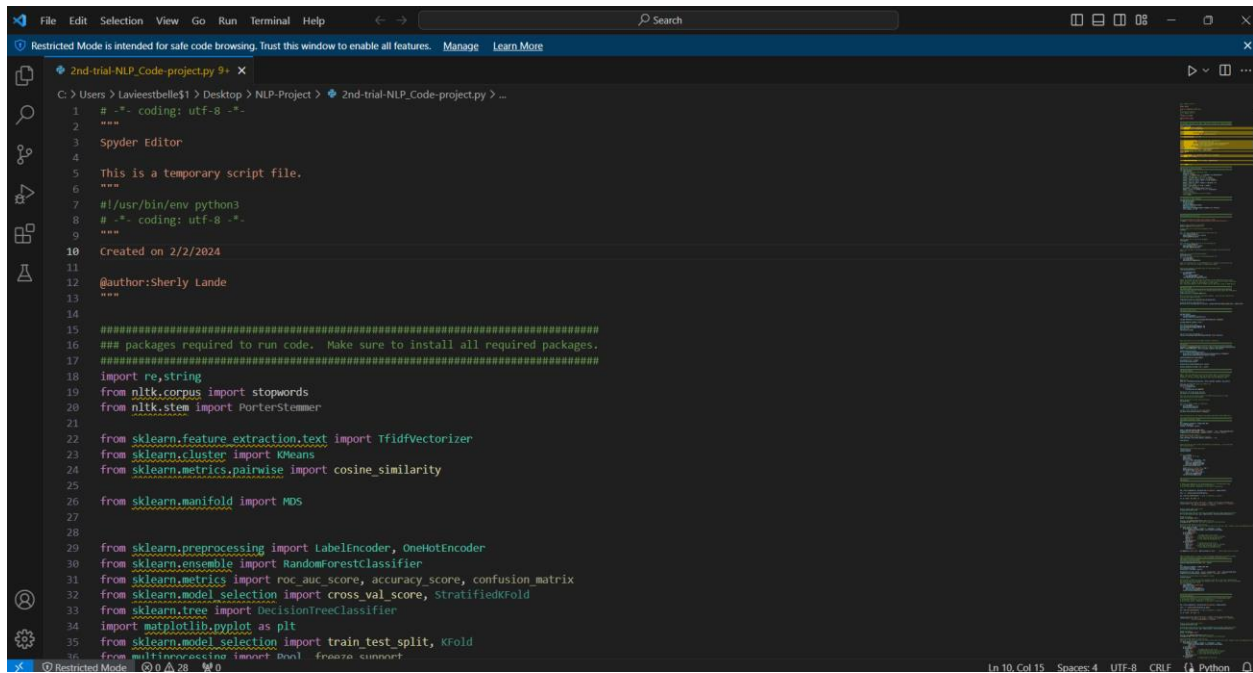
## References: Data

- Baines, Tom, Guillaume Zoppi, Leon Bowen, Thomas P. Shalvey, Silvia Mariotti, Ken Durose, and Jonathan D. Major. 2018. "Incorporation of CdSe Layers into CdTe Thin Film Solar Cells." *Solar Energy Materials and Solar Cells* 180: 196–204. <https://doi.org/10.1016/j.solmat.2018.03.010>.
- Bothwell, Alexandra M., Jennifer A. Drayton, and James R. Sites. 2020. "Performance Analysis of 0.4–1.2- $\mu\text{m}$  CdTe Solar Cells." *IEEE Journal of Photovoltaics* 10(1): 259–266.
- Dharmadasa, I. M., A. E. Alam, A. A. Ojo, and O. K. Echendu. 2019. "Scientific Complications and Controversies Noted in the Field of CdS/CdTe Thin Film Solar Cells and the Way Forward for Further Development." *Journal of Materials Science. Materials in Electronics* 30(23): 20330–44. <https://doi.org/10.1007/s10854-019-02422-6>.
- Devkota, Suman, Kwame Asiedu Owusu Nyako, Brendan Kuzior, Victor Karpov, Daniel G. Georgiev, Frank Li, Pedro Cortes, and Vamsi Borra. 2022. "Threshold Switching in CdTe Photovoltaics." *ECS Transactions* 109(1): 3–10. <https://doi.org/10.1149/10901.0003ecst>.
- Feng, Lianghuan, Lili Wu, Zhi Lei, Wei Li, Yaping Cai, Wei Cai, Jingquan Zhang, Qiong Luo, Bing Li, and Jiagui Zheng. 2007. "Studies of Key Technologies for Large Area CdTe Thin Film Solar Cells." *Thin Solid Films* 515(15): 5792–97. <https://doi.org/10.1016/j.tsf.2006.12.122>.
- Hatton, Peter, Michael J. Watts, Ali Abbas, John M. Walls, Roger Smith, and Pooja Goddard. 2021. "Chlorine Activated Stacking Fault Removal Mechanism in Thin Film CdTe Solar Cells: The Missing Piece." *Nature Communications* 12(1): 4938. <https://doi.org/10.1038/s41467-021-25063-y>.
- Khattak, Yousaf Hameed, Faisal Baig, Bernabé Marí, Saira Beg, Syed Rizwan Gillani, and Tanveer Ahmed. 2018. "Effect of CdTe Back Surface Field on the Efficiency Enhancement of a CGS Based Thin Film Solar Cell." *Journal of Electronic Materials* 47(9): 5183–90. <https://doi.org/10.1007/s11664-018-6405-4>.
- Kim, Hyoungseok, Kyoungsoon Cha, Vasilis M. Fthenakis, Parikhith Sinha, and Tak Hur. 2014. "Life Cycle Assessment of Cadmium Telluride Photovoltaic (CdTe PV) Systems." *Solar Energy* 103: 78–88. <https://doi.org/10.1016/j.solener.2014.02.008>.
- Kranz, Lukas, Christina Gretener, Julian Perrenoud, Rafael Schmitt, Fabian Pianezzi, Fabio La Mattina, Patrick Blösch, et al. 2013. "Doping of Polycrystalline CdTe for High-Efficiency Solar Cells on Flexible Metal Foil." *Nature Communications* 4(1): 2306. <https://doi.org/10.1038/ncomms3306>.
- Morales-Acevedo, Arturo. 2006. "Thin Film CdS/CdTe Solar Cells: Research Perspectives." *Solar Energy* 80(6): 675–81. <https://doi.org/10.1016/j.solener.2005.10.008>.
- Rajput, Pramod, Yogesh Kumar Singh, G.N. Tiwari, O.S. Sastry, Santosh Dubey, and Kailash Pandey. 2018. "Life Cycle Assessment of the 3.2 kW Cadmium Telluride (CdTe)

- Photovoltaic System in Composite Climate of India." *Solar Energy* 159: 415–22. <https://doi.org/10.1016/j.solener.2017.10.087>.
- Savado, O. 1998. "Chemically and Electrochemically Deposited Thin Films for Solar Energy Materials." *Solar Energy Materials and Solar Cells* 52(3): 361–88. [https://doi.org/10.1016/S0927-0248\(97\)00247-X](https://doi.org/10.1016/S0927-0248(97)00247-X).
- Tao, Coby S., Jiechao Jiang, and Meng Tao. 2011. "Natural Resource Limitations to Terawatt-Scale Solar Cells." *Solar Energy Materials and Solar Cells* 95(12): 3176–80. <https://doi.org/10.1016/j.solmat.2011.06.013>.
- Todorov, Teodor, Oki Gunawan, S. Jay Chey, Thomas Goislard de Monsabert, Aparna Prabhakar, and David B. Mitzi. 2011. "Progress towards Marketable Earth-Abundant Chalcogenide Solar Cells." *Thin Solid Films* 519(21): 7378–81. <https://doi.org/10.1016/j.tsf.2010.12.225>.
- Wolden, Colin A., Ali Abbas, Jiaojiao Li, David R. Diercks, Daniel M. Meysing, Timothy R. Ohno, Joseph D. Beach, Teresa M. Barnes, and John M. Walls. 2016. "The Roles of ZnTe Buffer Layers on CdTe Solar Cell Performance." *Solar Energy Materials and Solar Cells* 147: 203–10. <https://doi.org/10.1016/j.solmat.2015.12.019>.
- Yanyi Sun, Katie Shanks, Hasan Baig, Wei Zhang, Xia Hao. 2019. "Integrated CdTe PV Glazing into Windows: Energy and Daylight Performance for Different Window-to-Wall Ratio." *Energy Procedia*. <https://doi.org/10.1016/j.egypro.2019.01.976>.
- Zhao, Yuan, Mathieu Boccard, Shi Liu, Jacob Becker, Xin-Hao Zhao, Calli M. Campbell, Ernesto Suarez, Maxwell B. Lassise, Zachary Holman, and Yong-Hang Zhang. 2016. "Monocrystalline CdTe Solar Cells with Open-Circuit Voltage over 1 V and Efficiency of 17%." *Nature Energy* 1(6): 16067. <https://doi.org/10.1038/nenergy.2016.67>.
- Zyoud, Ahed H., Doa' H. Abdelhadi, Mohamed H.S. Helal, Samer H. Zyoud, Heba Bsharat, Sohaib M. Abu-Alrob, Nordin Sabli, Naser Qamhie, Abdul Razack Hajamohideen, and Hikmat S. Hilal. 2019. "Enhancement of Electrochemically Deposited Pristine CdTe Film Electrode Photoelectrochemical Characteristics by Annealing Temperature and Cooling Rate." *Optik (Stuttgart)* 197: 163220. <https://doi.org/10.1016/j.ijleo.2019.163220>.

## Appendices

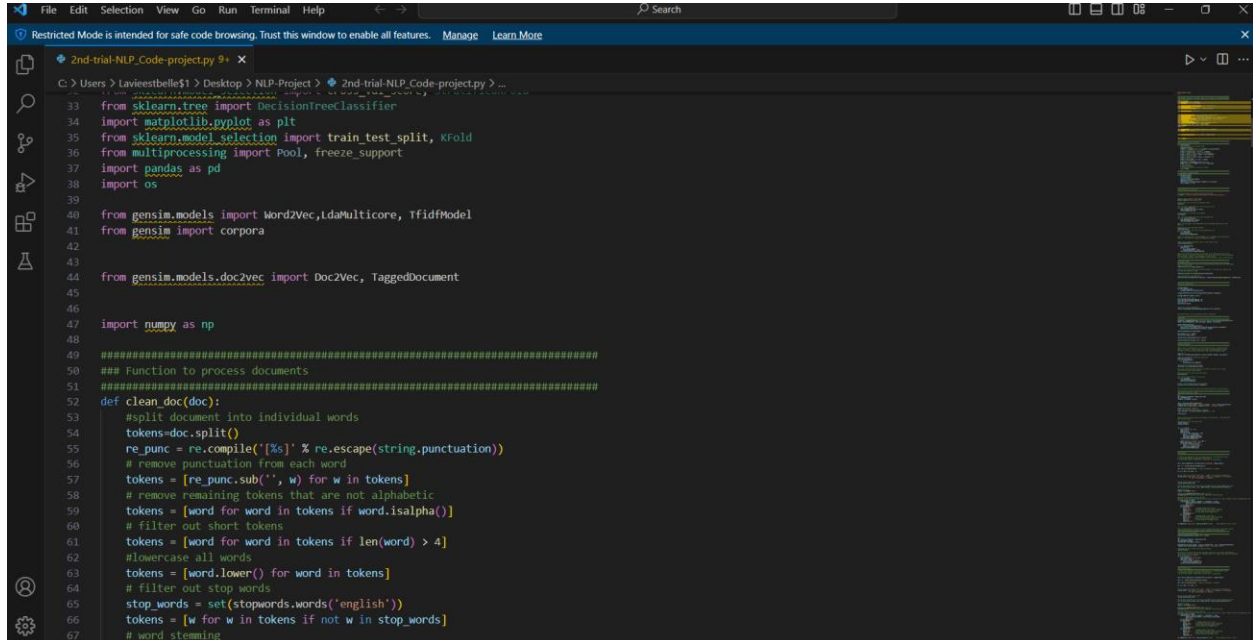
### Computational Methods for Semantic Analysis of Solar Cell Research Literature



```

1  # -*- coding: utf-8 -*-
2  """
3  Spyder Editor
4
5  This is a temporary script file.
6  """
7  #!/usr/bin/env python3
8  # -*- coding: utf-8 -*-
9  """
10 Created on 2/2/2024
11
12 @author: Sherly Lande
13 """
14
15 #####
16 ## packages required to run code. Make sure to install all required packages.
17 #####
18 import re, string
19 from nltk.corpus import stopwords
20 from nltk.stem import PorterStemmer
21
22 from sklearn.feature_extraction.text import TfidfVectorizer
23 from sklearn.cluster import KMeans
24 from sklearn.metrics.pairwise import cosine_similarity
25
26 from sklearn.manifold import MDS
27
28
29 from sklearn.preprocessing import LabelEncoder, OneHotEncoder
30 from sklearn.ensemble import RandomForestClassifier
31 from sklearn.metrics import roc_auc_score, accuracy_score, confusion_matrix
32 from sklearn.model_selection import cross_val_score, StratifiedKFold
33 from sklearn.tree import DecisionTreeClassifier
34 import matplotlib.pyplot as plt
35 from sklearn.model_selection import train_test_split, KFold
36 from multiprocessing import Pool, freeze_support

```



```

33 from sklearn.tree import DecisionTreeClassifier
34 import matplotlib.pyplot as plt
35 from sklearn.model_selection import train_test_split, KFold
36 from multiprocessing import Pool, freeze_support
37 import pandas as pd
38 import os
39
40 from gensim.models import Word2Vec, LdaMulticore, TfidfModel
41 from gensim import corpora
42
43
44 from gensim.models.doc2vec import Doc2Vec, TaggedDocument
45
46
47 import numpy as np
48
49 #####
50 ## Function to process documents
51 #####
52 def clean_doc(doc):
53     # split document into individual words
54     tokens = doc.split()
55     re_punc = re.compile('[%s]' % re.escape(string.punctuation))
56     # remove punctuation from each word
57     tokens = [re_punc.sub('', w) for w in tokens]
58     # remove remaining tokens that are not alphabetic
59     tokens = [word for word in tokens if word.isalpha()]
60     # filter out short tokens
61     tokens = [word for word in tokens if len(word) > 4]
62     # lowercase all words
63     tokens = [word.lower() for word in tokens]
64     # filter out stop words
65     stop_words = set(stopwords.words('english'))
66     tokens = [w for w in tokens if not w in stop_words]
67     # word stemming

```

```

File Edit Selection View Go Run Terminal Help
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More

2nd-trial-NLP_Code-project.py 9+ X
C:\Users\Lavieestbelle1\Desktop\NLP-Project> 2nd-trial-NLP_Code-project.py ...

46
47 import numpy as np
48
49 #####
50 ### Function to process documents
51 #####
52 def clean_doc(doc):
53     #split document into individual words
54     tokens=doc.split()
55     re_punc = re.compile('[%s]' % re.escape(string.punctuation))
56     # remove punctuation from each word
57     tokens = [re_punc.sub('', w) for w in tokens]
58     # remove remaining tokens that are not alphabetic
59     tokens = [word for word in tokens if word.isalpha()]
60     # filter out short tokens
61     tokens = [word for word in tokens if len(word) > 4]
62     #lowercase all words
63     tokens = [word.lower() for word in tokens]
64     # filter out stop words
65     stop_words = set(stopwords.words('english'))
66     tokens = [w for w in tokens if not w in stop_words]
67     # word stemming
68     ps=PorterStemmer()
69     # tokens=[ps.stem(word) for word in tokens]
70     return tokens
71
72 #####
73 # Functions to label encoding
74 #####
75 def One_Hot(variable):
76     LE=LabelEncoder()
77     LE.fit(variable)
78     Label1=LE.transform(variable)
79     OHE=OneHotEncoder()
80     labels=OHE.fit_transform(Label1.reshape(-1,1)).toarray()

```

```

File Edit Selection View Go Run Terminal Help
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More

2nd-trial-NLP_Code-project.py 9+ X
C:\Users\Lavieestbelle1\Desktop\NLP-Project> 2nd-trial-NLP_Code-project.py ...

72 #####
73 # Functions to label encoding
74 #####
75 def One_Hot(variable):
76     LE=LabelEncoder()
77     LE.fit(variable)
78     Label1=LE.transform(variable)
79     OHE=OneHotEncoder()
80     labels=OHE.fit_transform(Label1.reshape(-1,1)).toarray()
81     return labels, LE, OHE
82
83 #####
84
85 #####
86 ### Processing text into lists
87 #####
88
89 #set working Directory to where class corpus is saved.
90 os.chdir(r'C:\Users\Lavieestbelle1\Desktop\NLP-Project\word document\')
91
92
93 #read in class corpus csv into python
94 data=pd.read_csv('Class_Corpus.csv')
95
96 #create empty list to store text documents titles
97 titles=[]
98
99 #for loop which appends the OSI title to the titles list
100 for i in range(0,len(data)):
101     temp_text=data['OSI_Title'].iloc[i]
102     titles.append(temp_text)
103
104 #create empty list to store text documents
105 text_body=[]
106
107 #for loop which appends the text to the text body list

```

```

File Edit Selection View Go Run Terminal Help
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More

2nd-trial-NLP_Code-project.py 9+ X
C:\Users> Lavieestbelle$1 > Desktop > NLP-Project > 2nd-trial-NLP_Code-project.py > ...

88
89 #set working Directory to where class corpus is saved.
90 os.chdir(r"C:\Users\Lavieestbelle$1\Desktop\NLP-Project\word document/")
91
92
93 #read in class corpus csv into python
94 data=pd.read_csv('Class_Corpus.csv')
95
96 #create empty list to store text documents titles
97 titles=[]
98
99 #for loop which appends the DSI title to the titles list
100 for i in range(0,len(data)):
101     temp_text=data['DSI_title'].iloc[i]
102     titles.append(temp_text)
103
104 #create empty list to store text documents
105 text_body=[]
106
107 #for loop which appends the text to the text_body list
108 for i in range(0,len(data)):
109     temp_text=data['Text'].iloc[i]
110     text_body.append(temp_text)
111
112 #Note: the text_body is the unprocessed list of documents read directly form
113 #the csv.
114
115 #empty list to store processed documents
116 processed_text=[]
117 #for loop to process the text to the processed_text list
118 for i in text_body:
119     text=clean_doc(i)
120     processed_text.append(text)
121
122 #Note: the processed_text is the PROCESSED list of documents read directly form
123 #the csv.

```

```

File Edit Selection View Go Run Terminal Help
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More

2nd-trial-NLP_Code-project.py 9+ X
C:\Users> Lavieestbelle$1 > Desktop > NLP-Project > 2nd-trial-NLP_Code-project.py > ...

106
107 #for loop which appends the text to the text_body list
108 for i in range(0,len(data)):
109     temp_text=data['Text'].iloc[i]
110     text_body.append(temp_text)
111
112 #Note: the text_body is the unprocessed list of documents read directly form
113 #the csv.
114
115 #empty list to store processed documents
116 processed_text=[]
117 #for loop to process the text to the processed_text list
118 for i in text_body:
119     text=clean_doc(i)
120     processed_text.append(text)
121
122 #Note: the processed_text is the PROCESSED list of documents read directly form
123 #the csv. Note the list of words is separated by commas.
124
125
126 #stitch back together individual words to reform body of text
127 final_processed_text=[]
128
129 for i in processed_text:
130     temp_DSI=i[0]
131     for k in range(1,len(i)):
132         temp_DSI=temp_DSI+' '+i[k]
133     final_processed_text.append(temp_DSI)
134
135 #Note: We stitched the processed text together so the TFIDF vectorizer can work.
136 #Final section of code has 3 lists used. 2 of which are used for further processing.
137 #(1) text_body - unused, (2) processed_text (used in W2V),
138 #(3) final_processed_text (used in TFIDF), and (4) DSI titles (used in TFIDF Matrix)
139
140
141

```