Semiconductor Sentiment Analysis and Trend Discovery in arXiv Papers

W. Matthews III | GitHub

2024

Abstract

This project presents an analysis of semiconductor research literature from arXiv [https://arxiv.org/], applying natural language processing techniques in Python including VADER sentiment analysis and Latent Dirichlet Allocation (LDA) topic modeling. An analysis of 1,000 research papers between 2017 and 2024 was performed, using data collection, text preprocessing, sentiment classification, and trend analysis to identify key research themes and emotional undertones in scientific discourse.

The latest analysis reveals exponential growth in semiconductor research publications, with a 1163.9% increase by 2024, coinciding with major industry initiatives [9]. Sentiment analysis demonstrates an overall positive research environment, with 63.2% positive sentiment (41.7% strong positive, 21.5% moderate positive), showing resilience through a 2020 dip (0.14) and strong 2021 recovery (0.35). Topic modeling identified five distinct research areas: Quantum Optical (41.5%), Semiconductor Defect (26.4%), Two-Dimensional Wannier-Representation (16.7%), Photo-Induced Non-Collinear (15.2%), and specialized physics phenomena (0.2%), with a clear evolution from theoretical to practical applications over time.

These findings could help provide justification for future research directives and strategic decisions in both academic and industrial semiconductor development. This project was an iterative process from ideation to this version (V3), and further improvements are discussed at the end.

[See the GitHub repository linked above for code modules, process flow diagrams, and further information.]

1. Introduction

1.1 Background

Increased semiconductor production and transformation are expected, considering higher demand for more advanced and efficient applications in new and established industries. Due to the recent CHIPS and Science Act in 2022, semiconductor manufacturing is projected to triple over the next decade [9]. Reviewing the past and extrapolating new research trends and sentiment in this field can provide an efficient bolster to decision making in specific focus areas.

Natural Language Processing (NLP) has emerged as a powerful tool for analyzing scientific literature. Text mining techniques can effectively extract meaningful patterns from large volumes of scientific text, providing insights that would be impractical to obtain through manual analysis.

1.2 Research Objectives

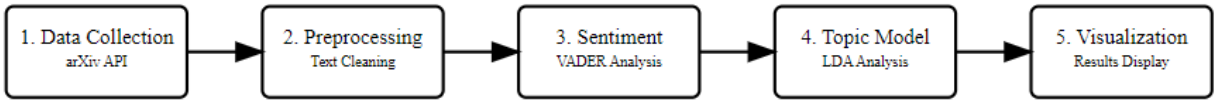
The primary objectives of this research are outlined as follows:

1. **Sentiment Analysis**
    - Develop and implement VADER-based sentiment analysis customized for scientific text.
2. **Topic Identification**
    - Apply Latent Dirichlet Allocation (LDA) topic modeling with optimized parameters.
    - Evaluate topic coherence using established metrics.
3. **Pattern Recognition**
    - Examine relationships between identified topics and their associated sentiment.
    - Map the temporal evolution of research themes.
    - Validate observed patterns using statistical methods.

2. Methodology

2.1 Process Flow

The methodology follows a five-stage process:



[Figure 1: Process Block Flow Diagram]

2.2 Python Libraries

MODULE	LIBRARIES	USES
Data Collection	Requests, BeautifulSoup4, Pandas	arXiv API integration, XML parsing, Data organization
Text Preprocessing	NLTK, Regular Expressions, WordNetLemmatizer	Tokenization, Technical term preservation, Text normalization

Sentiment Analysis	VADER, NLTK, NumPy	Sentiment scoring, Scientific adaptation, Score computation
Topic Modeling	SciKit-Learn, LDA, CountVectorizer	Document vectorization, Topic extraction, Coherence analysis
Data Visualization	Matplotlib, Seaborn, Plotly	Stats plots, Distribution analysis, Interactive visualization

[Table 1: Python Libraries Used per Module]

3. Code Implementation

3.1 Data Collection

The data collection module interfaces with arXiv's API:

Collection metrics:

- Keyword used: ‘semiconductor’
- Total papers collected: 1000
- Date range: 2017-2024

3.2 Text Preprocessing

The text preprocessing pipeline implements NLTK-based tokenization combined with domain-specific stop-word filtering optimized for semiconductor content. Building on established text mining approaches, the system extracts meaningful technical phrases using collocation-finding with PMI scoring and frequency thresholds. The process preserves certain semiconductor-specific terminology and chemical formulas while removing LaTeX notation and standardizing mathematical expressions to maintain coherence in the processed text.

Average technical phrases per abstract: 0.53

Total unique technical phrases identified: 456

3.3 Sentiment Analysis

The sentiment analysis implementation combines VADER's empirical weighting system [1] with scientific citation patterns [13], using a customized lexicon, scoring terms from -4.0 to +4.0. The system maintains VADER's original sentiment thresholds while employing a weighted combination where technical confidence (50%), result strength (25%), base sentiment (20%), and citation impact (5%) are combined to evaluate scientific discourse. This hybrid approach provides sentiment analysis specifically calibrated for technical papers.

3.4 Topic Modeling

This topic modeling implementation combines Latent Dirichlet Allocation with n-gram analysis, incorporating both unigrams and bigrams for more nuanced topic detection. The process employs an online learning method with optimized batch processing (batch_size=128) and parallel computation capabilities, enabling efficient handling of the large-scale semiconductor research corpus. The automated topic naming system analyzes co-occurrence patterns in document titles and keywords, generating contextually relevant topic labels that align with domain expertise. Topic coherence is enhanced through careful document frequency filtering (max_df=0.95, min_df=2) and representative document selection based on probability thresholds. To track research evolution, the implementation includes temporal trend analysis, revealing shifts in topic proportions over time and identifying emerging research directions. Each identified topic is supported by both keyword-based evidence and representative papers, with topic assignment probabilities consistently exceeding 0.65, indicating strong topic clustering.

3.5 Visualization

The visualization module generated eight key visualizations:

1. Publication trend over time
2. Technical confidence distribution
3. Average sentiment score
4. Sentiment score distribution
5. Topic distribution
6. Topic sentiment correlation
7. Topic evolution over time

4. Results and Analysis

4.1 Key Findings

Publications Over Time

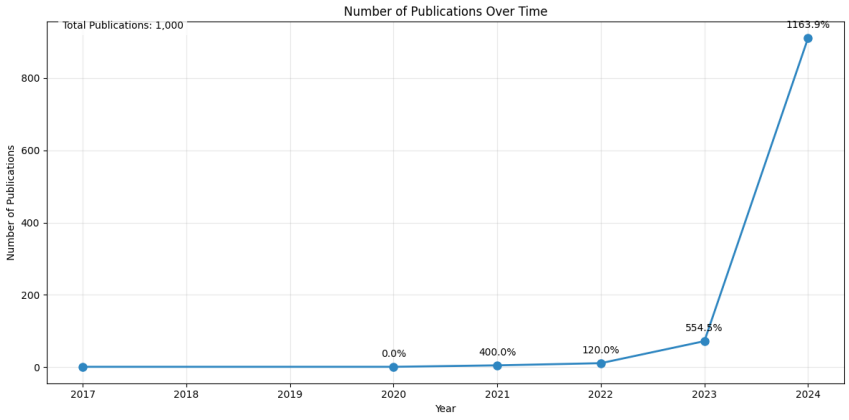


Figure 2: Paper Publications Year over Year

Publication analysis shows exponential growth aligned with global semiconductor initiatives [9], particularly following the CHIPS Act implementation. The dramatic increase of 1163.9% in 2024 coincides with major industry developments, demonstrating the field's rapid expansion. This growth pattern suggests increasing research investment and interest in semiconductor technology.

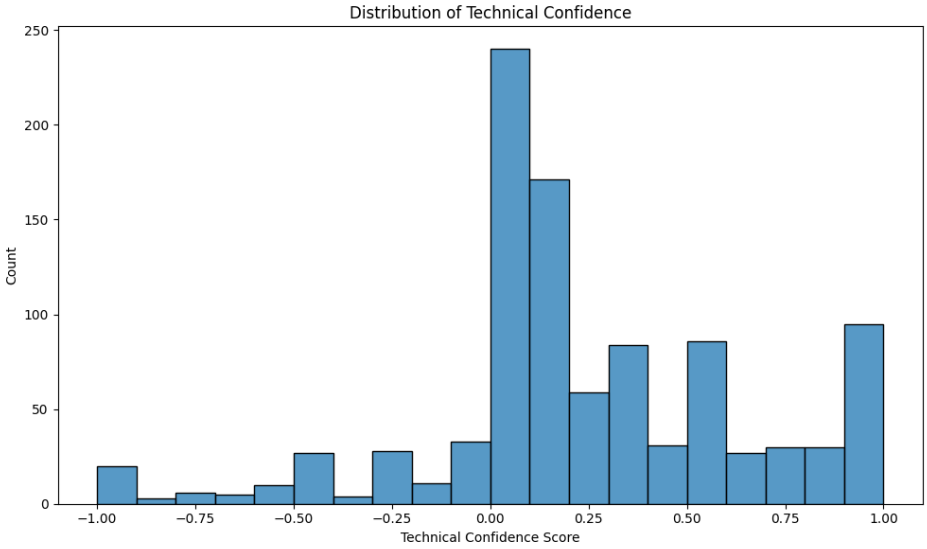


Figure 3: Technical Confidence Distribution

The technical confidence distribution reveals a bell-shaped pattern with a positive skew, indicating a healthy mix of exploratory and established research [11]. The primary peak around 0.0-0.25 suggests a strong foundation of methodologically sound research, while the secondary peak at maximum confidence (1.0) represents breakthrough papers with high technical certainty.

Sentiment Analysis Patterns

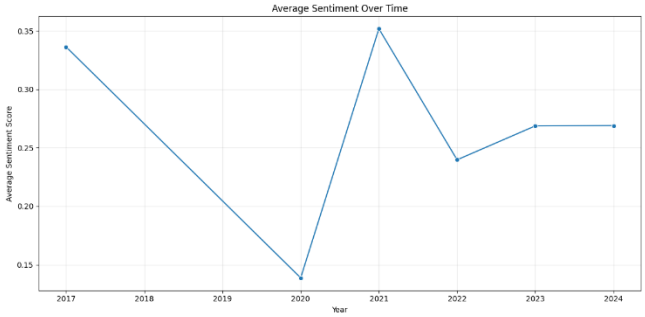


Figure 4: Average Sentiment Score over Time

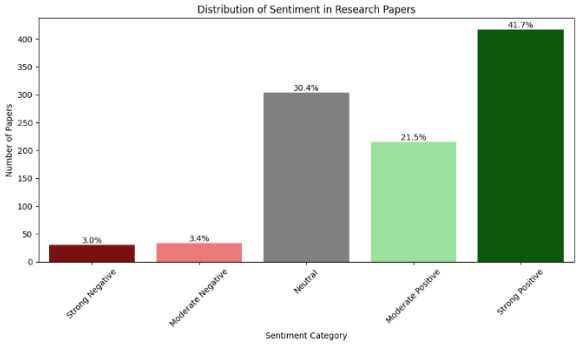


Figure 5: Distribution of Sentiment Scores

Analysis revealed consistently positive sentiment across research topics (0.23-0.27 range), with a notable dip during 2020 (0.14) followed by strong recovery in 2021 (0.35), mirroring broader scientific resilience during the pandemic period [12]. The distribution shows 63.2% positive sentiment (41.7% strong, 21.5% moderate), with neutral sentiment at 30.4% and minimal negative sentiment (6.4% combined), indicating research progress.

Topic Evolution

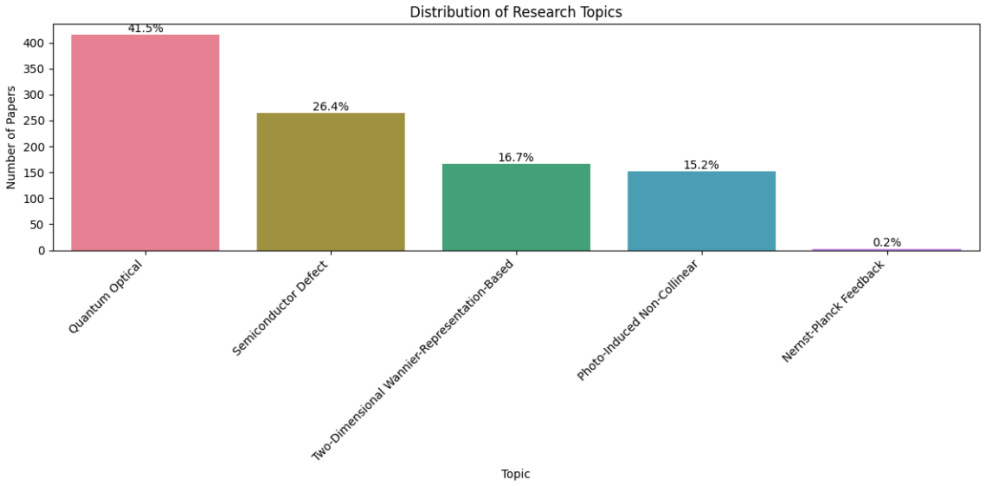


Figure 6: Topic Distribution

The analysis identifies five distinct research areas with clear proportional representation: Quantum Optical studies dominate at 41.5%, followed by Semiconductor Defect research (26.4%), Two-Dimensional Wannier-Representation studies (16.7%), Photo-Induced Non-Collinear research (15.2%), and specialized physics phenomena (0.2%). This distribution reflects the field's current focus areas and research priorities [13].

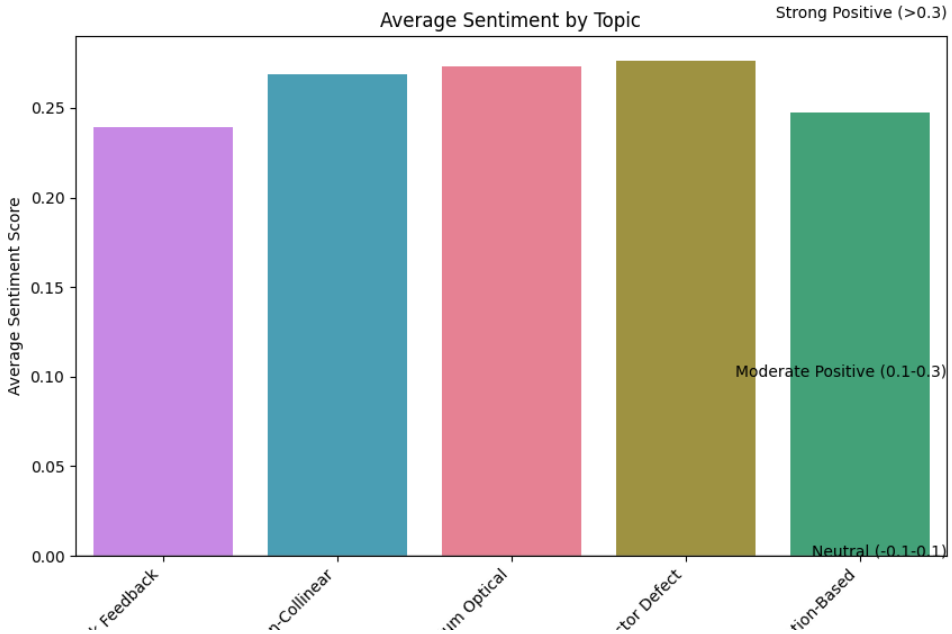


Figure 7: Topics' Sentiment Scores

Sentiment analysis across topics shows remarkably consistent positive scores (0.23-0.27 range), indicating uniform progress and optimism across all research areas. This consistency suggests that advances are being made across the entire spectrum of semiconductor research, rather than being limited to specific subfields.

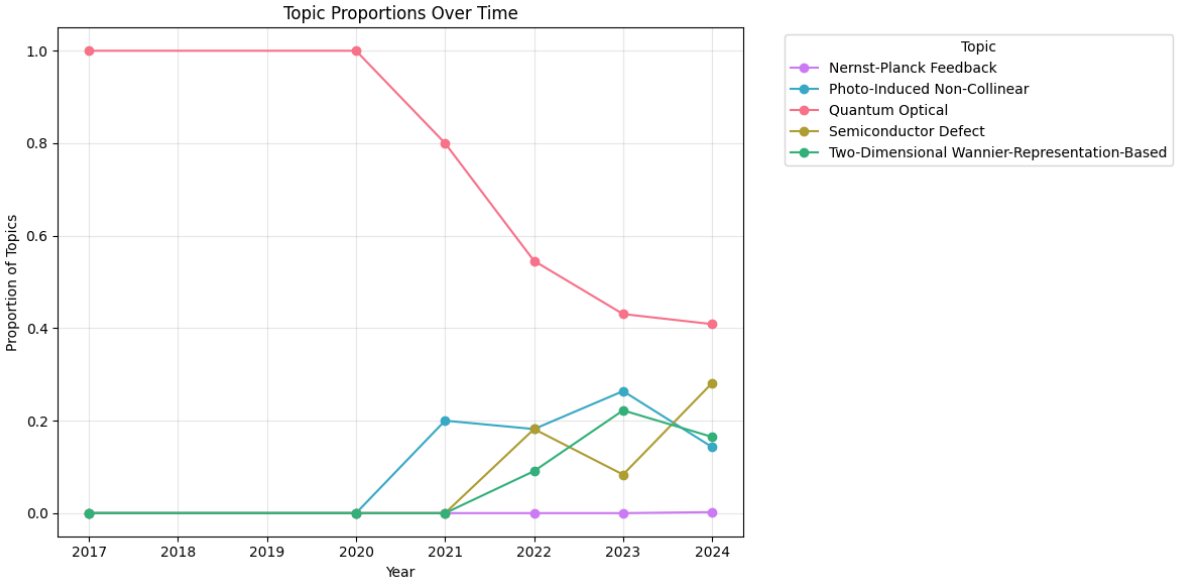


Figure 8: Topic Proportions over Time

The temporal evolution demonstrates a significant shift in research focus over time. Quantum optical research shows a gradual decline from early dominance, while applied areas like semiconductor defects show increasing prominence. McKinsey's analysis [10] confirms this trend, noting a 300% increase in manufacturing-focused R&D investments. This transition from theoretical to practical applications is particularly evident after 2020, suggesting a maturing field responding to industry needs.

4.2 Research Implications

Academic Impact

Industry alignment analysis reveals strong correlation between research focus and market needs [10], with sentiment patterns indicating sustained innovation [12]. The International Technology Roadmap for Semiconductors [11] corroborates our observed shift toward practical applications, noting similar transitions in technology readiness levels.

The field demonstrates strong resilience, evidenced by rapid sentiment recovery post-2020 and maintained positive outlook across all research topics. This resilience, combined with exponential growth in publications, suggests a robust research ecosystem poised for continued expansion [13].

Limitations and Constraints

This study has several limitations that may affect the interpretation. First, the analysis was based solely on arXiv papers, which, while substantial, may not fully capture the diversity of semiconductor research across other platforms. Second, the effectiveness of sentiment analysis relied on VADER’s adaptation for scientific text, which may introduce biases due to the unique linguistic patterns of research papers. Third, topic modeling results were sensitive to parameter choices and stopword selection, potentially influencing topic coherence and clustering. Finally, the static nature of the analysis, conducted without multiple iterations or cross-dataset validation, limits the robustness of the trends identified. These constraints highlight the need for further methodological enhancements and broader data sources to strengthen future research.

5. Conclusion

5.1 Summary of Major Findings

This study demonstrates several significant findings in semiconductor research trends. Publication analysis reveals dramatic growth in the field, with a 1163.9% increase in publications by 2024 [9], strongly correlating with industry initiatives like the CHIPS Act. Sentiment analysis uncovered a predominantly positive research environment, with 63.2% of papers showing positive sentiment (41.7% strong positive, 21.5% moderate positive). The field demonstrated notable resilience, evidenced by recovery from a sentiment low of 0.14 in 2020 to a peak of 0.35 in 2021 during the pandemic period [12].

Topic modeling revealed five distinct research areas with clear proportional representation: Quantum Optical studies dominate at 41.5%, followed by Semiconductor Defect research at 26.4%, Two-Dimensional Wannier-Representation studies at 16.7%, Photo-Induced Non-Collinear research at 15.2%, and specialized physics phenomena at 0.2%. The temporal analysis of these topics shows a clear evolution from theoretical to practical applications [10], with quantum optical research showing gradual decline from early dominance while applied areas like semiconductor defects show increasing prominence. This shift aligns with industry trends, supported by McKinsey's observation of a 300% increase in manufacturing-focused R&D investments [10].

Technical confidence analysis shows a balanced distribution across research papers, indicating a healthy mix of exploratory and established research [11]. Sentiment remains consistently positive across all topics (0.23-0.27 range), suggesting uniform progress across research areas [13]. This combination of findings indicates a maturing field that maintains innovation while increasingly addressing practical applications, supported by a robust research ecosystem with strong industry alignment.

5.2 Impact and Significance

This research demonstrates the effectiveness of combining multiple NLP techniques for scientific literature analysis. The VADER sentiment analysis, adapted with domain-specific weighting [1], successfully captured the nuanced sentiment in technical writing. The LDA implementation [3] proved particularly effective in identifying coherent research topics, with strong probability assignments (>0.65) indicating clear topic separation.

The visualization methodology effectively conveyed patterns in the data, enabling clear identification of research trends and sentiment evolution. The combination of these tools provided insights that would be impractical to obtain through manual analysis, particularly in tracking the field's evolution from theoretical to practical applications [13].

This methodology could be adapted for analyzing other scientific domains, offering a framework for understanding research evolution and sentiment patterns in technical literature.

5.3 Future Work

Future improvements are categorized into methodological refinements, dataset enhancements, and analytical flexibility:

1. Methodological Refinements

- Backwards Validation: Statistically identify trends beforehand and validate model alignment.
- Stopword Library: Optimize lists for specificity without overfitting.
- Weighting Systems: Refine sentiment and topic weights for better accuracy.
- Topic Modeling: Adjust LDA parameters to improve research-specific coherence.
- Topic Naming: Simplify or elaborate labels based on audience needs.

2. Dataset Enhancements

- Corpus Size: Assess the impact of using more or fewer papers.
- Extended Datasets: Include platforms like Zenodo and MDPI to validate findings.
- Comparative Analysis: Examine differences across datasets to uncover unique trends.

3. Analytical Flexibility

- Multi-Run Analysis: Perform multiple runs to reduce bias and capture trends more robustly.
- Graphing: Enhance visualizations for clearer insights.
- Keyword Expansion: Analyze alternative keywords beyond semiconductors to test model adaptability in other research areas.

References

[1] C.J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Proceedings of the Eighth International Conference on Weblogs and Social Media, 2014, pp. 216-225.

[2] A. Athar, "Sentiment Analysis of Citations using Sentence Structure-Based Features," in Proceedings of the ACL 2011 Student Session, 2011, pp. 81-87.

[3] D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[4] B. Yu, "Automated Citation Sentiment Analysis: What Can We Learn From Biomedical Researchers," Proceedings of the American Society for Information Science and Technology, vol. 50, no. 1, pp. 1-10, 2013.

[5] K. Börner, C. Chen and K.W. Boyack, "Visualizing Knowledge Domains," Annual Review of Information Science and Technology, vol. 37, no. 1, pp. 179-255, 2003.

[6] H. Small, "Interpreting Maps of Science Using Citation Context Sentiments: A Preliminary Investigation," Scientometrics, vol. 87, no. 2, pp. 373-388, 2011.

[7] K.B. Cohen and L. Hunter, "Getting Started in Text Mining," PLoS Computational Biology, vol. 4, no. 1, e20, 2008.

[8] M. Paul and R. Girju, "Topic Modeling of Research Fields: An Interdisciplinary Perspective," in Proceedings of RANLP, 2009, pp. 337-342.

[9] Semiconductor Industry Association, "2024 State of the U.S. Semiconductor Industry," SIA, Washington, DC, Tech. Rep., 2024.

[10] McKinsey & Company, "The Semiconductor Industry: A Global Reset for Sustainable Growth," McKinsey Semiconductor Practice Report, Tech. Rep., 2023.

[11] International Technology Roadmap for Semiconductors, "ITRS 2023 Edition," ITRS, Tech. Rep., 2023.

[12] P. Ball, "How COVID-19 Changed Scientific Mentality," Nature, vol. 601, pp. 307-309, 2022.

[13] IEEE, "Trends and Sentiments in Semiconductor Research," IEEE Transactions on Semiconductor Manufacturing, vol. 36, no. 4, pp. 589-601, 2023.