Embedding-based Trend Analysis in NeurIPS Publications

Kosei Uemura
Department of Computer Science, University of Toronto
k.uemura@mail.utoronto.ca

March 14, 2025

1 Introduction

The rapid advancement of machine learning and artificial intelligence has led to an exponential growth in research output over the past few decades. Conferences such as NeurIPS serve as a premier venue for presenting groundbreaking work in these fields. In this study, we focus on a dataset comprising NeurIPS papers spanning from 1987 to 2023. Our primary objective is to explore how the characteristics of these publications—such as the number of papers, abstract lengths, and title properties—have evolved over time.

The motivation for this analysis stems from a dual purpose: to refine our data science skills in data acquisition, cleaning, and visualization, and to gain insights into the historical trends and shifts in research focus within the NeurIPS community. By formulating the research question "How have the characteristics of NeurIPS papers evolved over the years?" we aim to establish a foundation for further investigations, which will later incorporate advanced embedding analysis for topic extraction and trend mapping.

2 Methods

2.1 Data Acquisition

The dataset was obtained by web scraping the NeurIPS proceedings website. To accommodate structural differences across different years, two distinct scraping routines were developed. For the years 1987 to 2021, a general scraping function was implemented using Python's `requests` and `BeautifulSoup` libraries, which enabled retrieval of HTML content and extraction of paper titles and links. For the years 2022 and 2023, a specialized routine was crafted to account for modifications in the website's structure, ensuring accurate extraction of paper details. The extracted metadata, including titles and abstracts, were then consolidated into CSV files for subsequent analysis.

2.2 Data Wrangling and Cleaning

Following data collection, the dataset was merged and cleaned to ensure consistency and reliability. Records with missing data or with extremely short abstracts (i.e., abstracts containing fewer than five words, noted as ”Abstract Unavailable”) were filtered out to maintain data quality. Additionally, feature engineering was performed by computing supplementary attributes such as the word counts of titles and abstracts. These computed metrics were instrumental in identifying potential anomalies and ensuring that only relevant papers were retained for analysis.

2.3 Exploratory Data Analysis and Visualization

An exploratory analysis was conducted to uncover trends and patterns within the dataset by count-base. Overall paper trends were examined by analyzing the annual publication counts, revealing growth patterns over time. Textual characteristics of the papers were explored by computing summary statistics for abstract and title lengths, offering insights into changes in writing style and content density. In addition, publication-quality figures were generated using Python libraries such as `matplotlib` and `seaborn`. These visualizations, including line plots that depict the evolution of publication counts and median abstract lengths over time, provided valuable context and supported the formulation of hypotheses regarding shifts in research focus and publication practices.

3 Preliminary Results

We focus on three main aspects: the total number of accepted papers over the years, the average title length, and the median abstract length.

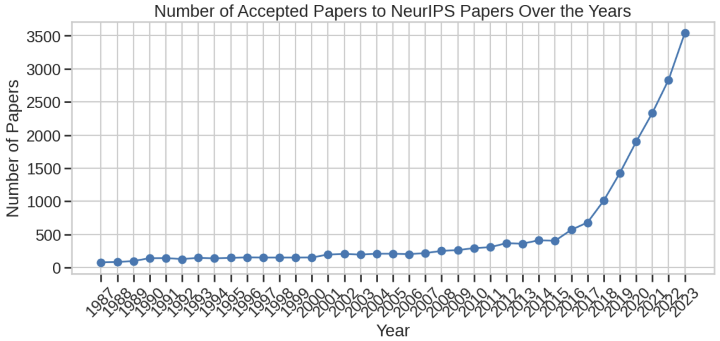

Figure 1: Number of accepted NeurIPS papers from 1987 to 2023. The substantial increase in the last decade underscores the growing research interest and participation in the conference.

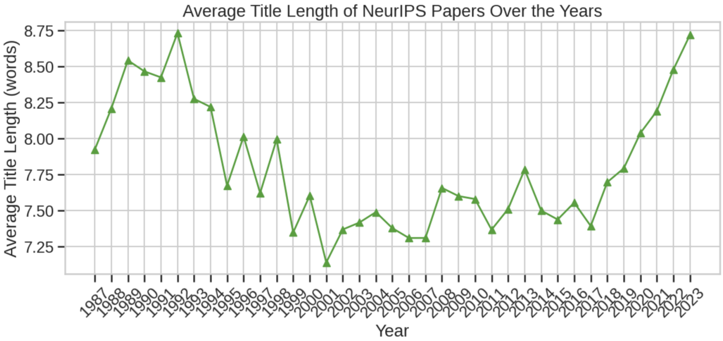

Figure 2: Average title length (in words) of NeurIPS papers over time. Although there are minor fluctuations, the average length remains relatively stable.

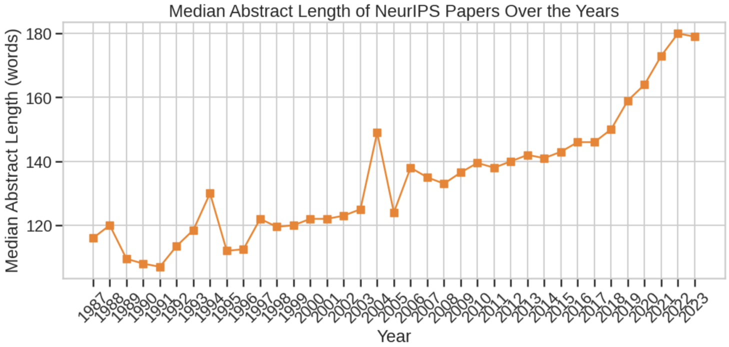

Figure 3: Median abstract length (in words) of NeurIPS papers from 1987 to 2023. A generally increasing trend is evident, possibly indicating growing complexity and detail in research descriptions.

Figure 1 shows the number of accepted papers from 1987 to 2023. A substantial rise in the number of papers is observed, particularly in the last decade, suggesting both increased research output and potentially broader participation within the NeurIPS community. The acceleration in publications may be attributed to growing interest in machine learning and artificial intelligence, as well as the conference's evolving submission and review policies.

Figure 2 presents the average title length over the years. Although some fluctuations are visible, the overall trend remains relatively stable, hovering around 8 to 9 words per title.

Figure 3 highlights the median abstract length across the same time period. There is a general upward trend in abstract length, suggesting that authors may be providing more comprehensive summaries of their work. This observation may reflect the increasing complexity of research problems and the need to convey broader methodological details. However, it is also possible that changing norms and guidelines for abstract formatting have played a role.

Overall, these preliminary findings offer an initial glimpse into how NeurIPS publications have evolved. The subsequent analysis will dive deeper into the content of abstracts and titles using advanced techniques such as embedding-based clustering and topic modeling. This deeper investigation will aim to uncover finer-grained insights into the evolving research directions within the NeurIPS community.

4 Summary

The preliminary exploration of NeurIPS papers from 1987 to 2023 indicates a pronounced increase in the overall volume of accepted papers, particularly in recent years. Concurrently, the median length of abstracts has exhibited a steady rise, suggesting a trend toward more comprehensive overviews of research findings.

These observations partially address our primary research question on how the characteristics of NeurIPS publications have evolved over time. While the upward trajectory in paper count highlights the growing popularity of machine learning and related fields, the expanded abstract lengths may reflect an increased complexity in research methodology and a broader set of experimental results.

4.1 Plan for the Final Project

For the final project, we will extend this analysis by integrating an embedding-based pipeline to perform in-depth trend and topic modeling. Specifically, we plan to:

1. **Embed Texts with SPECTER2:** Use the `allenai/specter2_base` model to obtain high-dimensional embeddings for titles and abstracts. This step will capture semantic nuances that are not apparent through simple word counts or frequency-based methods.

2. **Dimensionality Reduction and Clustering:** Apply UMAP for dimensionality reduction, followed by hierarchical clustering using HDBSCAN. We will further consolidate clusters if they exceed a predefined threshold in number, and assign outliers via a k-nearest neighbors approach. This procedure will group papers into meaningful topic clusters.

3. **Topic Extraction and Growth Analysis:** Derive important keywords from each cluster using n-gram extraction and a class-based TF-IDF (c-TF-IDF) approach. We will then calculate the growth rate of each topic cluster over time to identify emerging or declining research areas.

4. **Interactive Visualization:** Generate an interactive map of the embedded papers using `datamapplot`. This visualization will enable users to explore the distribution of topics, hover over individual points for additional details, and search for specific keywords or papers of interest.

By implementing these steps, we aim to provide a comprehensive portrait of how NeurIPS research topics have evolved and to highlight salient thematic clusters driving current innovations in machine learning and artificial intelligence. This deeper analysis will offer a richer understanding of the temporal and thematic trends within the conference.

5 Appendix

The Appendix provides detailed statistical summaries of the title and abstract lengths for NeurIPS papers across the years. Table 1 presents the title length statistics (mean, median, minimum, and maximum number of words) for each year, offering insight into the evolution of titling conventions over time. Similarly, Table 2 summarizes the abstract lengths after filtering, highlighting trends in the depth and detail of paper abstracts. These supplementary tables complement the main analysis by providing a granular view of how the textual characteristics of publications have changed over the years.

Year	**Mean**	**Median**	**Min**	**Max**
1987	7.92208	7	3	19
1988	8.20732	8	3	16
1989	8.54082	8	2	18
1990	8.46429	8	2	18
1991	8.42254	8	3	18
1992	8.73016	9	2	18
1993	8.27397	8	2	20
1994	8.21739	8	2	16
1995	7.67123	7	2	16
1996	8.01316	8	2	19
1997	7.62	7	2	18
1998	7.99333	8	2	19
1999	7.34667	7	2	18
2000	7.60265	7	1	16
2001	7.13846	7	2	15
2002	7.36765	7	1	18
2003	7.41538	7	2	15
2004	7.48792	7	2	16
2005	7.37681	7	1	14
2006	7.30846	7	2	16
2007	7.30876	7	2	15
2008	7.656	8	1	16
2009	7.59924	8	2	15
2010	7.57877	7	2	15
2011	7.36601	7	1	20
2012	7.50954	7	2	16
2013	7.78333	8	2	16
2014	7.49878	7	2	15
2015	7.43672	7	2	18
2016	7.5536	7	3	16
2017	7.39175	7	2	15
2018	7.69574	7	2	17
2019	7.79202	8	1	19
2020	8.03583	8	1	21
2021	8.1898	8	2	19
2022	8.47671	8	2	20
2023	8.71667	9	2	23

Table 1: Title Length Statistics per Year (in words)

Year	**Mean**	**Median**	**Min**	**Max**
1987	179.727	116	35	1257
1988	152.378	120	23	828
1989	167.480	109.5	34	1383
1990	129.793	108	26	474
1991	134.528	107	32	1113
1992	127.484	113.5	6	361
1993	129.027	118.5	6	332
1994	172.471	130	48	1053
1995	140.438	112	17	946
1996	194.408	112.5	34	3384
1997	166.020	122	13	1138
1998	178.407	119.5	47	1559
1999	168.453	120	6	1627
2000	173.629	122	39	1069
2001	202.087	122	8	3450
2002	166.691	123	44	1912
2003	155.385	125	9	1378
2004	944.565	149	11	4723
2005	191.203	124	32	3409
2006	222.353	138	58	3116
2007	160.977	135	19	2541
2008	149.196	133	44	2749
2009	147.546	136.5	35	1581
2010	143.476	139.5	45	298
2011	144.186	138	22	516
2012	144.886	140	5	300
2013	146.150	142	41	292
2014	149.939	141	36	303
2015	147.596	143	58	293
2016	151.316	146	59	307
2017	152.268	146	40	313
2018	154.403	150	38	298
2019	162.481	159	55	308
2020	168.194	164	23	317
2021	176.142	173	48	339
2022	181.272	180	17	364
2023	182.145	179	29	369

Table 2: Abstract Length Statistics per Year (in words, after filtering)