Ofir Beck

NLP final project - bibliometric analysis:

Introduction:

NLP has rapidly evolved in recent years, from rule-based systems to large-scale, transformer-based architectures, fundamentally reshaping how machines understand human language.

The academic publications on NLP should reflect the rapid evolution of its methodologies and applications in recent years, thus providing an interesting topic for a comprehensive bibliometric analysis of NLP papers from 2015 to 2025, using Scopus as the primary data source.

I combined bibliometric methods with modern NLP techniques to extract key insights about the development and structure of this growing field.

Data collection:

I have gathered 1000 articles from Scopus for each year between 2015-2025, then merged them into one CSV of 11000 total articles. Retrieving the same number of articles from each year allowed for a less biased analysis of the research trends over the 10 years.

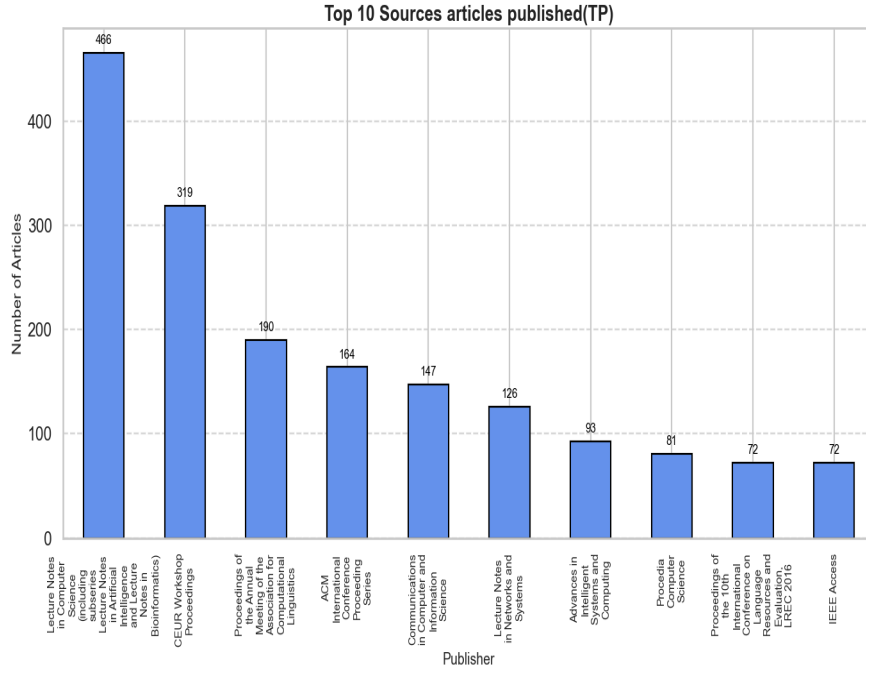
Performance analysis:

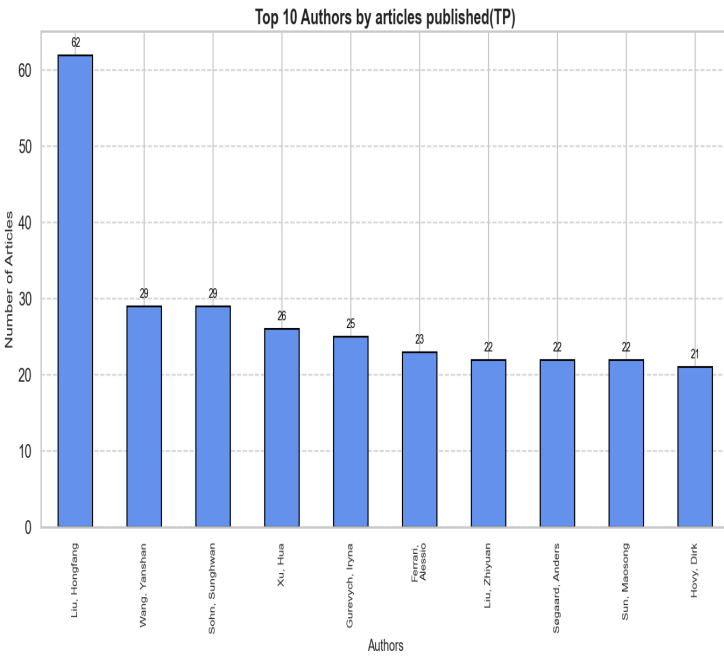
To fully analyze the sources and authors of those articles, I have conducted productivity, impact measures, as well as collaboration indicators.

Productivity measures:

I have calculated TP for each of the authors and sources from the 11000 articles and found the highest ones. Due to having a limited number of articles that are likely not the only ones published, it is a limited measurement.

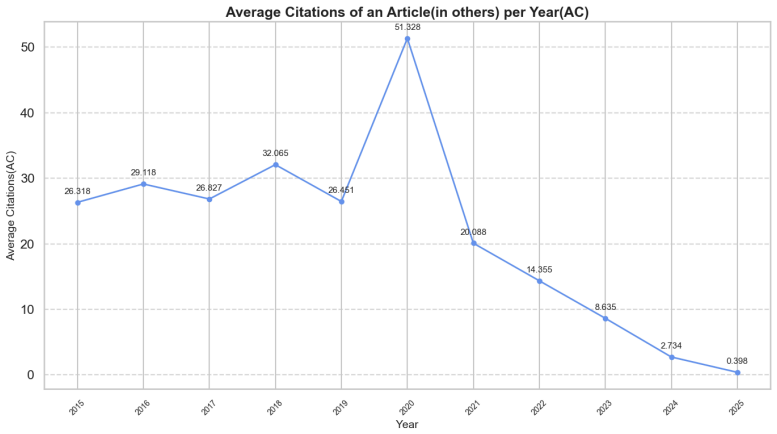
Since Scopus does provide updated citation count for each article, measuring the impact of articles, authors, and sources is much easier to do accurately - no need to have all the articles that referenced them in our dataframe.

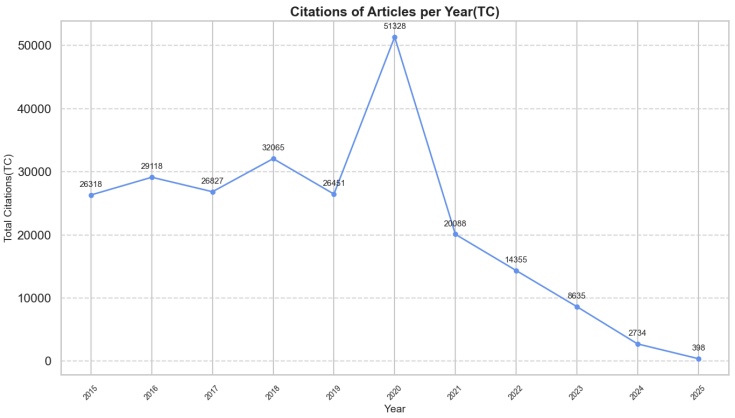




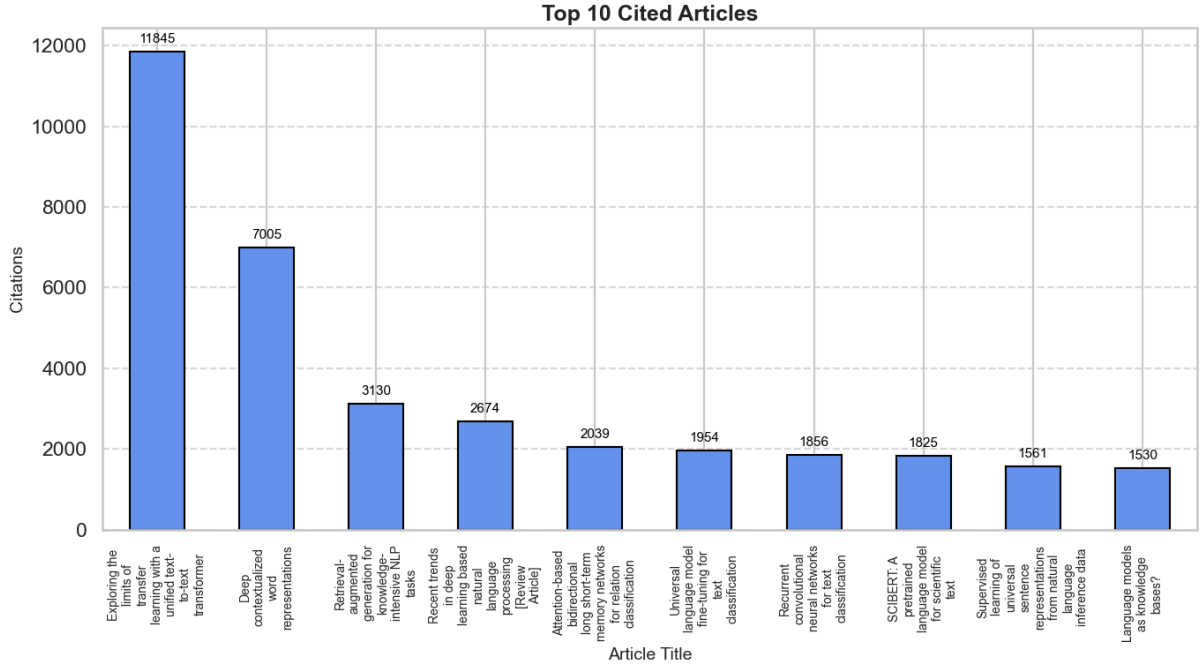
Impact Measures:

Firstly, I have measured both TC and AC of the published articles over the years(since there are 1000 articles from each year, AC = TC/1000):



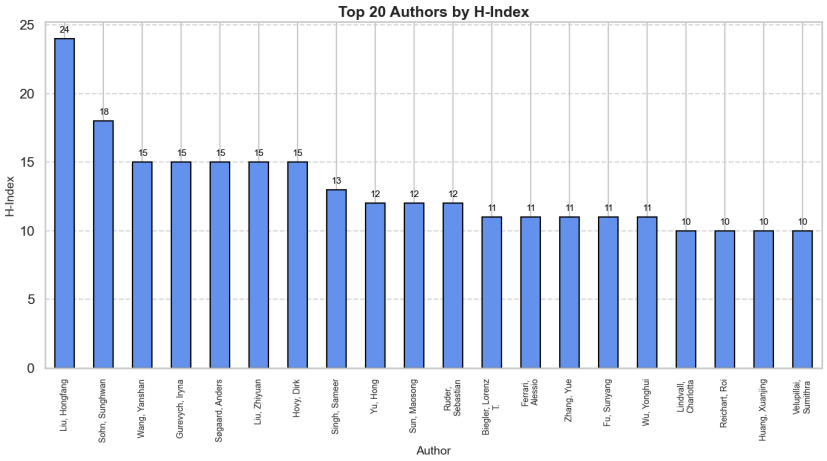


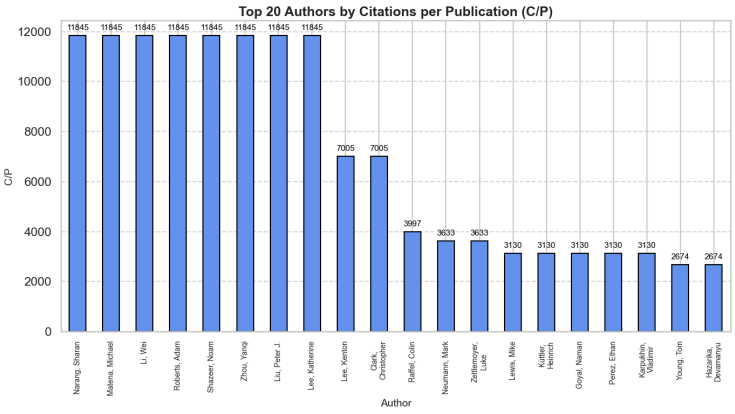
Those graphs indicate that the most referenced articles are from 2020, which was already into the BERT revolution, and the time when GPT-3 and T5 were released. One of the main reasons for such a peak may be due to the inclusion of the **"Exploring the limits of transfer learning with a unified text-to-text transformer"** in the dataframe, which is one of the most influential NLP papers of that decade, introducing the T5. As can be seen in the most cited graph:



The paper **“Deep Contextualized Word Representations”**(2018) was the 2nd most cited, which introduced ELMo - one of the first learning contextual word embeddings, where the representation of a word depends on its sentence context.

To compare both the productivity and the impact of authors, I have calculated the h-index of each one of them within the papers gathered. A researcher has an h-index of h if they have published h papers that have each been cited at least h times, reflecting consistent productivity and impact across multiple works. In addition, I have calculated the Citations per Publication (C/P) - an average number of citations each of a researcher's publications receives, highlighting the typical impact or influence per paper, regardless of how many papers they’ve published, thus mainly resembling impact on the field.



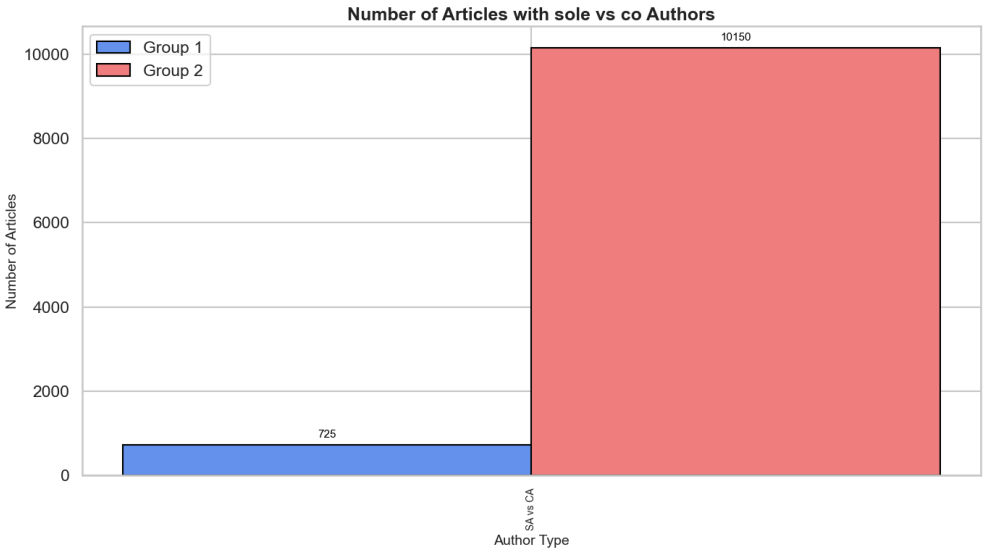


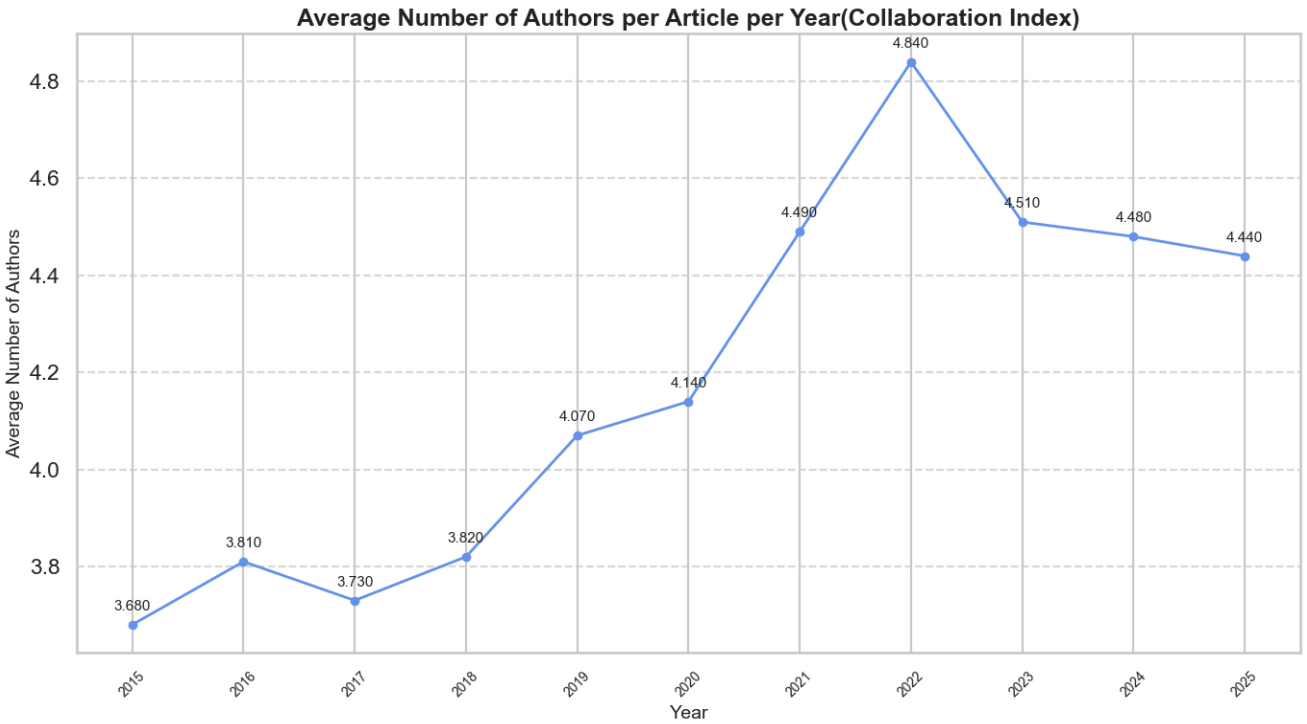
Due to a limited number of papers gathered, the true h-index and C/P of each author can’t be calculated exactly since there is a need for the true total amount of papers published by them, but it does give an insight.

Liu. Hongfang has been found to have the highest h-index here - 24, he is also with the highest TP of 62, which suggests consistent productivity and impact in this field through those years.
The authors with the highest C/P(both the 11845 and 7005) are the authors of the most cited articles in the df stated earlier, since all of them published only those papers in the df, their C/P are the high citation numbers of those papers / 1, with these papers those authors did make a huge impact on the field. We could also see that there isn’t necessarily a tight connection between those two metrics.

Collaboration indicators:

It can be seen that the NLP research field is a more collaborative one, with more than 90% of the papers conducted by more than one researcher. The collaboration index tells us how collaborative the field really is, seeing a constant increase of researchers per paper between 2017-2022, which may be due to the increase of complexity in that research, and then a slight decrease in recent years.

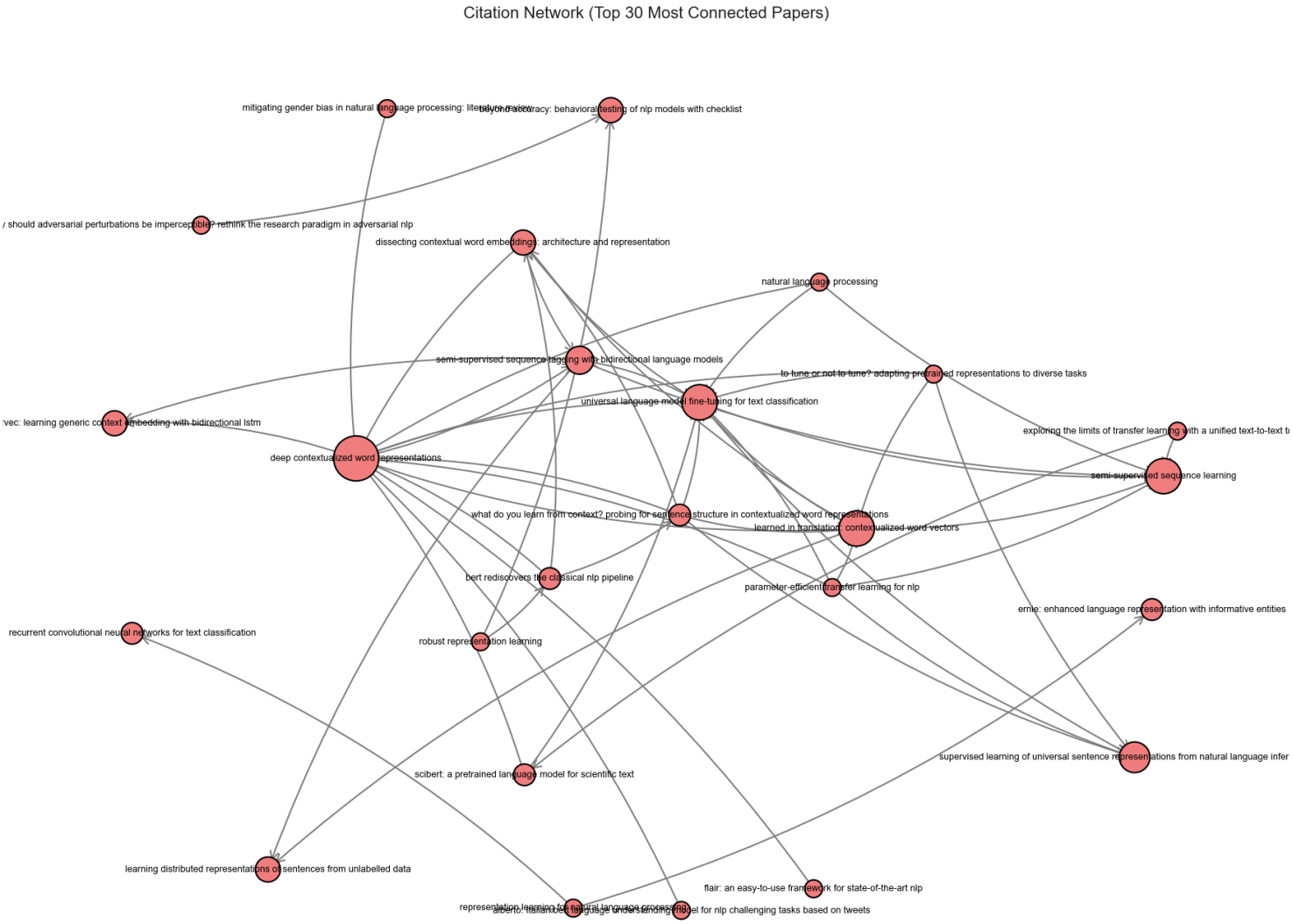




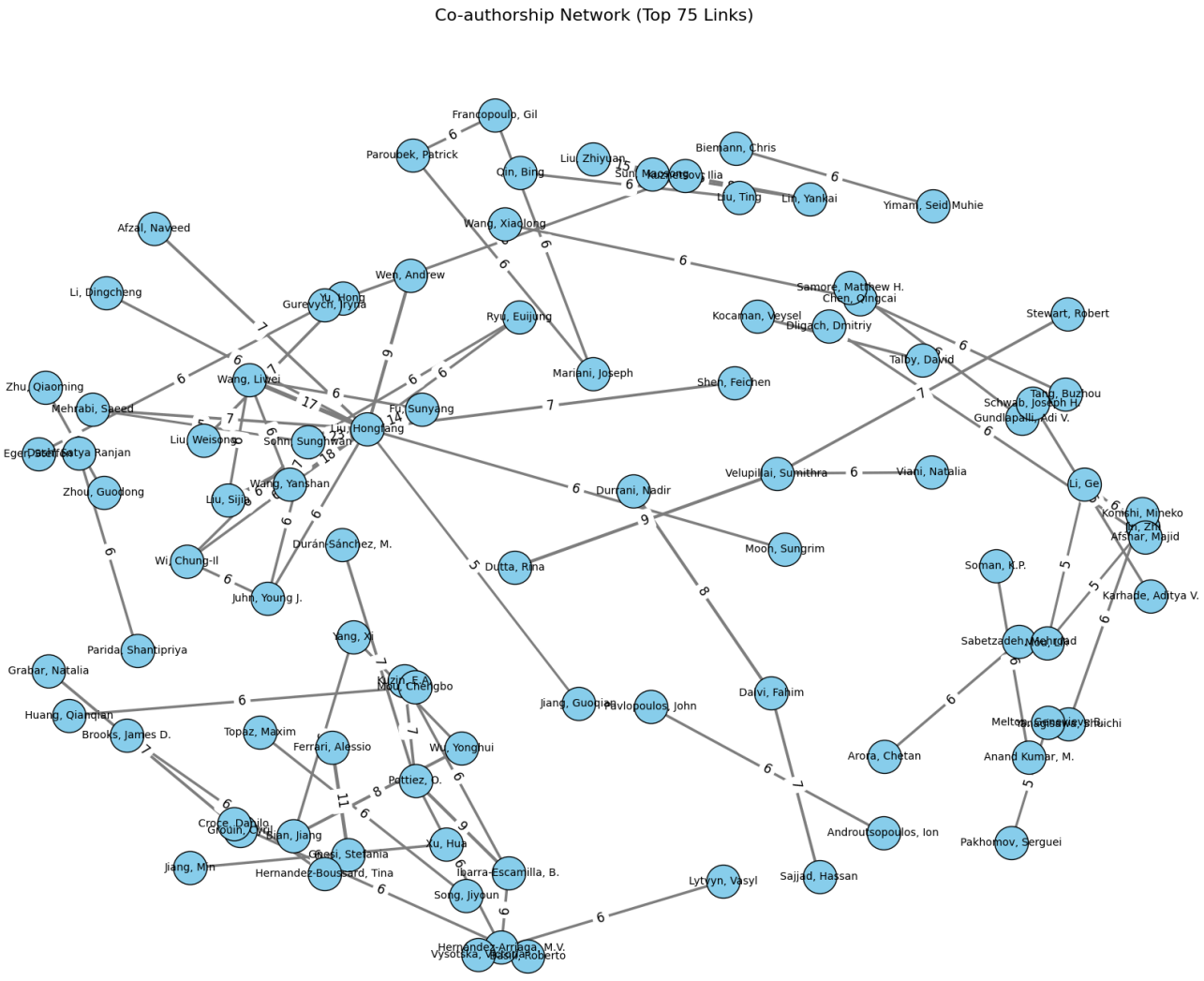
Science Mapping:

With network-based techniques, I have conducted a few analyses:

Citation analysis: I have linked the papers by direct citation, resulting in finding the most influential papers within this group. Two papers are connected in this network if one references the other(the one that is referencing is pointing to the referenced); the bigger the node is, the more it is referenced.
Supersingly the most connected and influencing paper within this group is **“Deep Contextualized Word Representations”** rather than **“Exploring the limits of transfer learning with a unified text-to-text transformer”** which is with the most citations overall, likely due to the fact of the former being older, and thus increasing the chance of fetching ones that reference it.

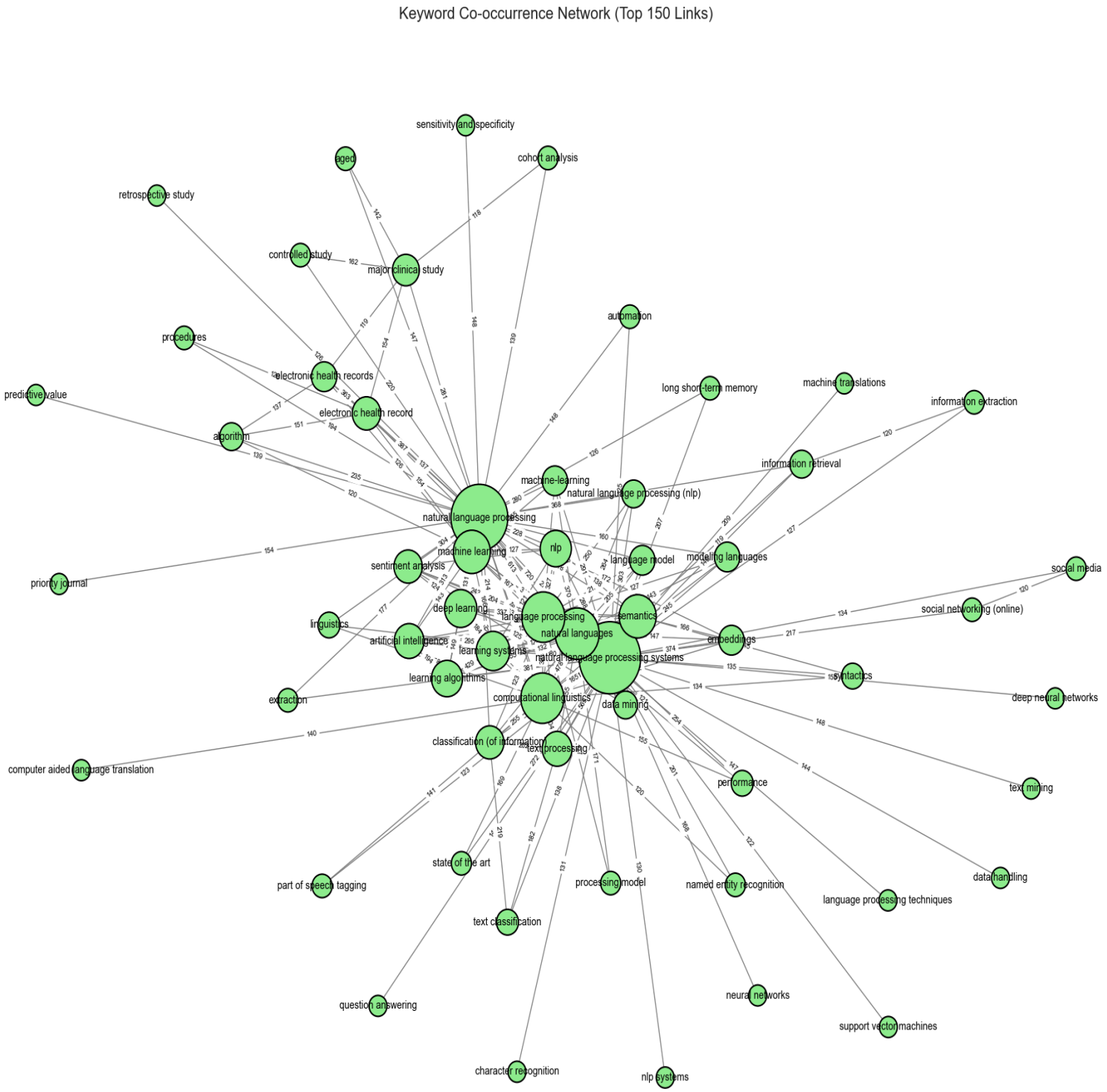


Co-authorship analysis: Maped collaboration networks among authors, showing the most collaborated authors within this article group, while the weights - the number of times those authors collaborated.
It is possible to point out Liu. Hongfang, which is connected to another author the most amount of times, shows not only its consistent productivity in the field, but also its collaboration skills.



Keyword analysis: Built networks of keywords co-occurring in the same article. We can see certain clusters forming, like:

- Electronic health record, algorithm, medical records, major clinical study - suggesting a focus on clinical data analysis using NLP.
- Social media and social networking are initiating the use of NLP in social media data analysis.
- Information retrieval and information extraction.
- Sentiment analysis together with DL, learning algorithms, and extraction.
- NLP is the most central keyword, along with deep learning.



I was particularly interested in showing growing trends in NLP research over the years, to better understand how the research in this field has progressed over the years. Though the keywords provided for each article were quite narrow, I have managed to find which keywords were more frequent in certain years than in others: Some clear stand-out ones are “COVID-19” rising in 2022, LLMs and chatbots dominating 2025 research.

	Place 1	Place 2	Place 3	Place 4	Place 5	Place 6	Place 7	Place 8
2015	classification	bioinformatics	genetics	artificial intelligence	electronic medical record	semantic web	procedures	text mining
2016	knowledge management	word sense disambiguation	ontology	optimization	nonlinear programming	linguistics	information analysis	semantic web
2017	education	software engineering	information analysis	convolution	neural networks	information retrieval systems	convolutional neural network	social networking (online)
2018	nlp (natural language processing)	recurrent neural networks	network architecture	long short-term memory	text processing	convolutional neural network	convolution	named entity recognition
2019	electronic health records	algorithm	medical informatics	electronic health record	major clinical study	search engines	sensitivity and specificity	word embedding
2020	question answering	computer aided language translation	convolutional neural networks	state of the art	deep neural networks	text classification	fine tuning	machine translations
2021	transfer learning	pre-training	pipelines	sensitivity and specificity	retrospective studies	predictive value	cohort analysis	word embedding
2022	covid-19	down-stream	controlled study	major clinical study	retrospective study	forecasting	predictive value	processing model
2023	e-learning	performance	transformers	processing model	character recognition	transformer	language processing	language model
2024	network security	contrastive learning	adversarial machine learning	economic and social effects	language processing techniques	data handling	machine translation	benchmarking
2025	emotion recognition	large language models	large datasets	knowledge graphs	large language model	teaching	diagnosis	chatbots

Conclusion:

This bibliometric analysis of NLP research from 2015 to 2025 reveals a field in rapid evolution, driven first by innovations in contextual embeddings (ELMo, 2018) and then by large-scale transfer-learning models (T5, GPT-3) peaking in 2020. By sampling 1000 articles per year, I ensured a balanced view of productivity (TP) and impact (TC, AC), and observed that high-output authors like Liu Hongfang also tend to be collaboration hubs, while single-publication authors (like the authors of the top-cited T5 and ELMo papers) achieve exceptional C/P ratios.

Collaboration metrics underscore NLP's cooperative nature - over 90% of papers involve multiple authors, with an increasing collaboration index from 2017-2022, mirroring the growing complexity of model development and data annotation workflows. Science-mapping networks further highlight how clinical NLP, social-media analysis, and sentiment-driven deep learning form a distinct yet linked field, with many unique uses. Keyword-based analysis has shown the trends of NLP research over the years.

Despite these insights, this approach is limited by the 1000-papers-per-year sampling, causing missing published articles of each author and citation-window bias favoring older publications. Future work could expand the data amount and sources, normalize citations by publication age, and apply dynamic network analyses to capture year-by-year shifts in subfields. Overall, this study demonstrates that a hybrid bibliometric NLP framework can effectively chart both the structural and conceptual development of a rapidly evolving field like NLP.