MGTF495 Final Report
Text Mining and NLP Application in Fluence Energy Analysis

Group Alpha_1: Yuankai Tao[*], Xinping Qu, Boning Xing, Yimin Luo

GitHub Repository

Homework MGTF495

Problem Statement

Our project focuses on leveraging text mining and Natural Language Processing (NLP) to analyze Fluence Energy's strategic expansion, entity relationships, sentiment regarding ESG (Environmental, Social, Governance), and the impacts of IRA policy incentives. Plus, we also try to conduct a light fine-tuning on bert-based model to construct a sentiment classifier based on the datasets loaded from hugging face link here, though this is still at the early stage but we would like to share what we have done so far and the intermediate result we have already got on Github.

Contributions

- Identifies relationships and financial interactions with other entities.
- Highlights the company's focus on ESG issues.
- Evaluates the impact of IRA (Inflation Reduction Act) policies on business operations and strategic planning.

Requirements and Dependencies

- 10-K annual financial reports spanning three years.
- External databases and dependencies for NLP and data analysis (Spacy, nltk, Transformers, Bert_based, FinBert, EnvironBert).

[*] Contact information: yut033@ucsd.edu, Address: San Diego, CA, 92122

Future Work

- Expand sentiment analysis to more granular ESG subcategories.
- Integrate real-time market data to correlate textual insights with financial performance indicators.
- Develop automated reporting dashboards to visualize relationships, sentiment trends, and policy impacts dynamically.
- Enhance the robustness of relationship extraction models through additional training on sector-specific documents.

Questions Statement

1. How is Fluence's expansion process? For example, get to know the changes in installed capacity and compare them with industry changes.
2. Perform Named Entity Recognition (NER) on three years of 10-K, and use the library tool to find out company entities in the 10-K and analyze their relationships with Fluence Energy, Inc.
3. Basic analysis includes: LDA topic analysis, word frequency analysis and sentiment analysis.
4. Count the frequency and proportion of ESG keywords that appear in the 10-K each year to see which issues have risen/fallen most significantly. For ESG-related paragraphs or sentences, classify the sentiment or opinion to examine whether the company is actively promoting or passively complying.
5. Explore the impact of IRA policy factors on the company. Check the frequency of mentions of IRA and related keywords in 10-K in the past three years, and compare it with ESG(Q4)and conduct time series analysis to see if there are changes year by year.

1. How is Fluence's expansion process?

The change in energy storage installed capacity is a key indicator for judging the expansion of energy storage companies. Therefore, the goal is to extract the annual installed capacity data of the company and global and finally draw a trend comparison chart.

The blue bars represent the Gross Global Pipeline (GW), which refers to the total global project pipeline, not limited to the company. The orange, green, and red bars represent Fluence's actual annual installed capacity. While the global energy project pipeline is expanding rapidly, Fluence is also showing steady year-over-year growth in both energy storage and renewable installations, as well as in its contracted backlog.

We tried a lot of keywords and finally found that using the unit of installed capacity as the keyword extraction is the most comprehensive and accurate. We use the Spacy library to perform simple natural language extraction, extract sentences related to the keyword (GW, MW, etc.) of this indicator, and make an initial observation.

Below are some of the extraction results we have obtained. We have marked different categories in different colors for easier reading. Through observation, we have learned that energy storage

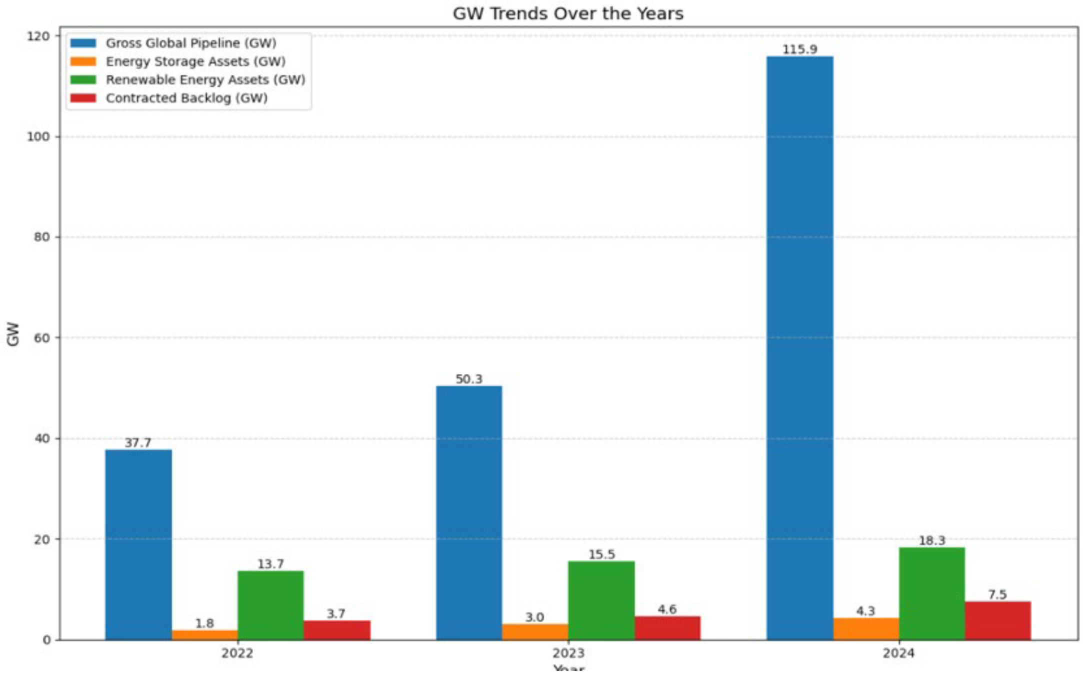


Figure 1: GW Trends

assets, renewable energy assets and contracted backlog are the contents we want to further extract and plot.

We encountered difficulties when using Spacy to extract and organize the data we need. We found that although the sentence structure of the 10-K text introducing the installed capacity is basically ”xx GW of xx”, there are exceptions, which makes Spacy unable to accurately identify the data and the corresponding category.

Therefore, we chose to use the BERT model in the free transformer library Hugging Face to complete the task. BERT stands for Bidirectional Encoder Representations from Transformers. It is a pre-trained deep learning model developed by Google in 2018 and is designed to understand the context of words in a sentence more accurately.

Each matched sentence is stored alongside the company pair. These relational sentences are then used as context for semantic analysis.

To interpret the nature of the relationship within each sentence, we used a transformer-based question answering (QA) model provided by Hugging Face's transformers library. The model receives a relational sentence as context and a question formatted as “What is the relationship between Fluence Energy Inc. and Company_XX?” It then returns a natural language answer extracted from the sentence.

This final step adds semantic depth to our pipeline. Rather than simply reporting co-occurrence, we now obtain a clear textual explanation of the relationship between the companies, which can be used for graph-based modeling or direct reporting.

There are 3 main relationships:

1. **Parent–Subsidiary Relationships**
   Fluence Energy GmbH (Germany),
   Fluence Energy Pty Ltd (Australia),
   Fluence Energy Singapore Pte. Ltd. (Singapore),
   Fluence Energy B.V. (Netherlands),

Fluence Energy AG (Switzerland), etc.

These entities are the subsidiaries of Fluence Energy across multiple countries.

2. **Shareholders and Investors**
   Several strategic stakeholders maintain significant ownership interests in Fluence Energy Inc. or its subsidiaries:

   - AES Grid Stability holds 28.5% of LLC interests in Fluence Energy LLC, which corresponds to 66.6% of the voting power in Fluence Energy Inc.
   - Siemens Industry Inc. previously held LLC interests but converted them in 2022 into Class A common stock of Fluence Energy Inc.
   - Qatar Holding LLC owns 34.2% of economic interest in Fluence Energy Inc., represented by 18,493,275 shares of Class A common stock.

   These ownership structures highlight the role of major multinational corporations and sovereign wealth funds in influencing Fluence's equity and corporate governance.

3. **Financial and Credit Relationships**

   - Fluence Energy Inc. also engages with financial institutions through formal credit arrangements: The company maintains an asset-based syndicated credit agreement (ABL Credit Agreement).
   - Barclays Bank PLC serves as the administrative agent for a $400 million revolving credit facility.

   During the extraction process, our initial filtering step yielded a total of 52 organization entities.

However, upon closer inspection, we observed a high degree of redundancy among these entities. This was due to the way organizations were referenced multiple times in slightly different formats throughout the 10-K document. For instance, the same company could appear with and without legal suffixes, or with formatting variations. Since spaCy treats each unique textual span as a distinct entity, such variations led to a repetition of the entity name.

To address this issue, we originally attempted to apply automated fuzzy matching using the `fuzzywuzzy` library to identify and merge similar entries. Unfortunately, the results were unsatisfactory, as the threshold-based similarity measures either collapsed distinct companies or failed to detect close variants due to subtle domain-specific differences.

As a result, we opted to perform manual duplication for this project. We reviewed the list of extracted entities and consolidated those that clearly referred to the same legal or operational entity. The final results presented in this report therefore reflect a clean and deduplicated set of company relationships, providing a more accurate and interpretable view of Fluence Energy Inc.'s corporate ecosystem.

2. Basic analysis including: LDA topic analysis, word frequency analysis and sentiment analysis

Step 1: Document Content Extraction

We extract text from a PDF file using `PDFMiner`, which processes each page and appends the extracted text to a list with `‘‘###PAGE_BREAK###’’` as a separator. It also handles potential errors. Moreover, we save the extracted text into a `.json` file so we can conveniently resume processing later without starting over.

Step 2: Natural Language Processing

Before conducting text analysis, we imported `nltk` and `spaCy` stopword libraries and merged them to obtain a more comprehensive stopword list. For example, in this section, the word “include” appears frequently and lacks specific meaning, so we added some domain-specific stopwords (e.g., finance and financial services industry terms) to the stopword library. In short, the stopword list must be well-prepared.

Step 3: Sentence and Paragraph Processing

We used regular expressions and other techniques to extract valid content. This includes, but is not limited to:

1. Remove ASCII non-printable characters, such as page numbers or PDF control characters.
2. Handle line breaks for sentence merging — as we’re consolidating broken lines into paragraphs, we want to make sure not to include headers.
3. Set a maximum of 500 WORDS PER PAGE — this avoids excessive length and ensures each page is treated as a valid paragraph.
4. Identify independent figure captions and irregular figure sections.
5. Remove empty content.
6. Remove newline characters.
7. Remove trailing spaces.

Word Cloud Map

From the word cloud analysis, although we were able to extract some information, the insights were quite limited, and the key points of the report remained unclear. Therefore, we visualized the top 30 most frequent words using a bar chart to gain a clearer understanding. The results are as follows:

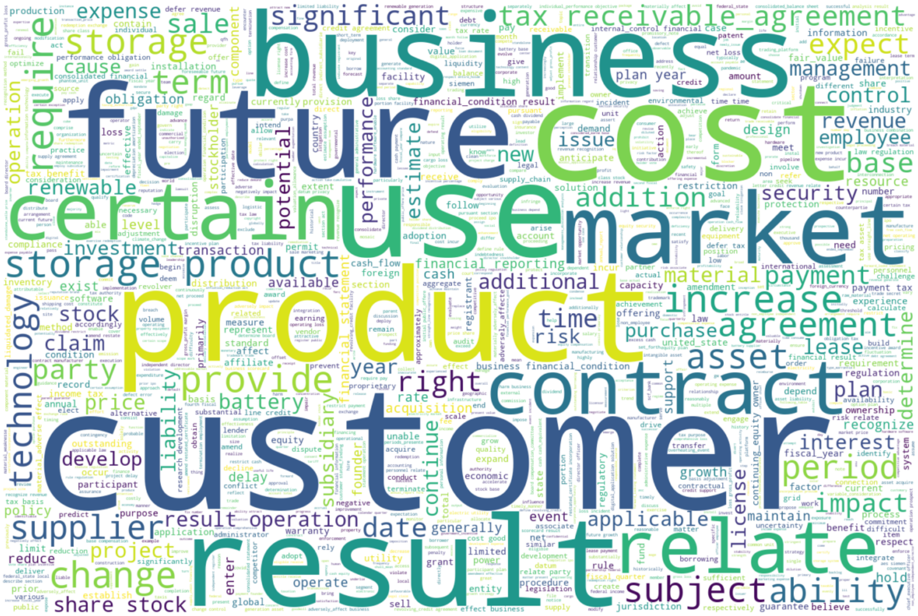


Figure 2: Word Cloud Example for 2022 10-K Report

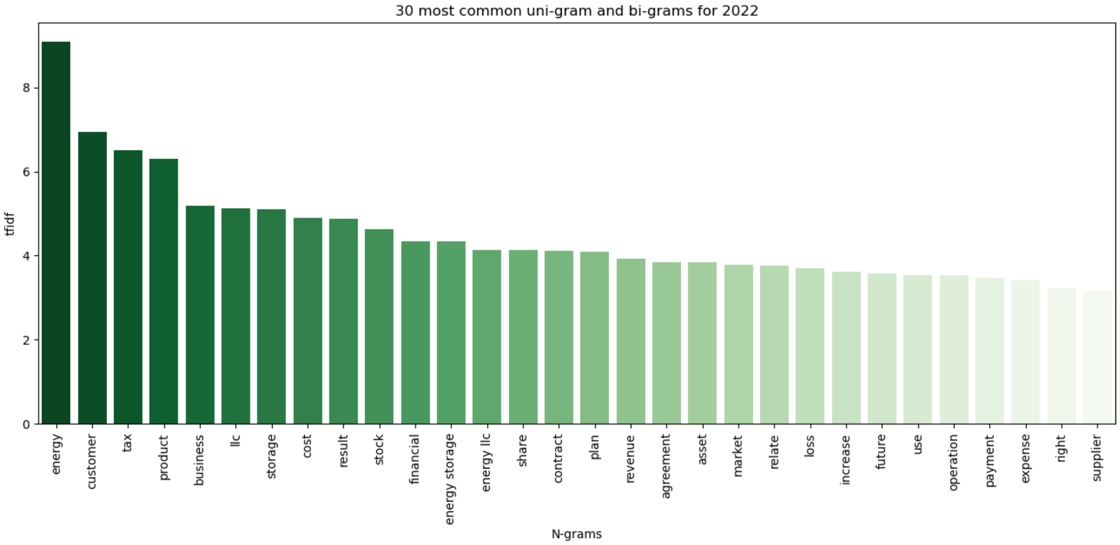


Figure 3: Top30 Words for 2022 10-K Report

2022 2024 Ngrams

We can clearly observe from the 2024-word frequency chart that the term ”energy storage” has jumped significantly in prominence—rising from 13th place in 2022 and 12th in 2023 to 6th place in 2024. This shift indirectly reflects an increase in the company's disclosure related to energy storage in its annual report. It also indicates that Fluence is placing greater strategic emphasis on the topic. Several factors may have contributed to this trend:

- Rising Market Demand: As global dependence on renewable energy continues to grow, energy storage technologies have become increasingly critical for grid stability and energy management. This rising importance likely prompted the company to devote more attention to the topic.

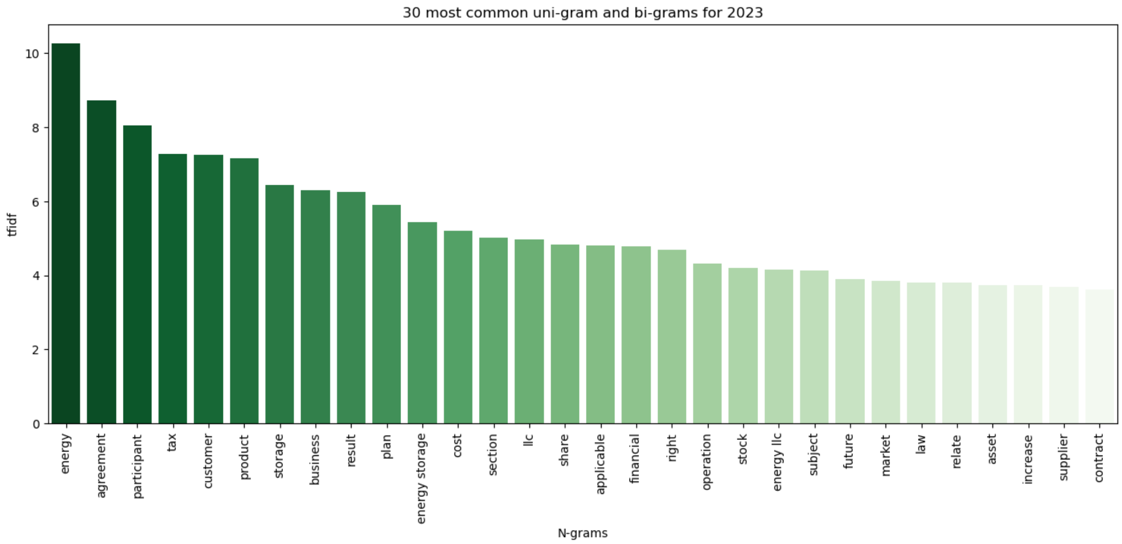


Figure 4: Top30 Words for 2023 10-K Report

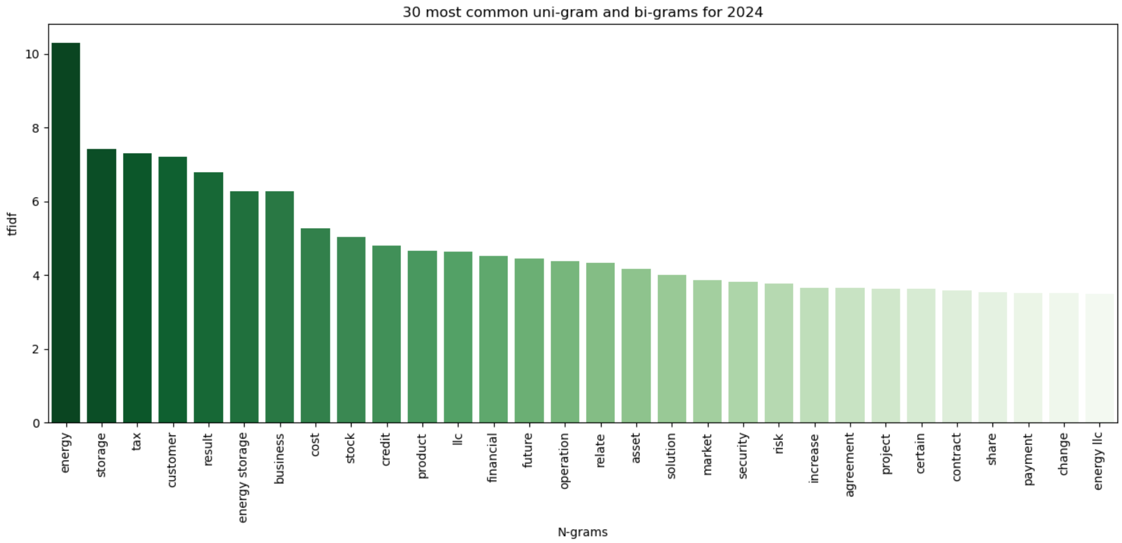


Figure 5: Top30 Words for 2024 10-K Report

- Policy and Regulatory Drivers: Governments around the world have introduced policies to support renewable energy and energy storage technologies—such as subsidies, tax incentives, and carbon emission regulations—which have encouraged companies like Fluence to enhance related disclosures.

- Technological Advancements and Cost Reduction: Ongoing improvements in energy storage technologies, coupled with declining costs, have made storage solutions more commercially viable. As a result, Fluence has been more inclined to highlight its progress and capabilities in this area.

LDA Topic Modeling

Additionally, we applied the Latent Dirichlet Allocation (LDA) topic model to perform unsupervised classification of the annual reports for each year. The primary objective was to uncover the underlying thematic structure of the reports by identifying distinct topics discussed within the texts.

Two key evaluation metrics were considered in model selection:

1. **Perplexity**: This metric reflects how well the model predicts unseen data, measured as the normalized log-likelihood on a held-out test set. Lower values indicate better generalization.

2. **Coherence Score**: This measures the semantic similarity between the high-probability words within each topic. Higher coherence indicates better interpretability of the topics.

In our analysis, we selected the model with the highest coherence score as the optimal configuration. Taking 2022 as an example, we found that the coherence score peaked at `0.4311549` when the number of topics `n=4`. Therefore, we consider dividing the 2022 annual report into four main thematic clusters to be the most appropriate choice.

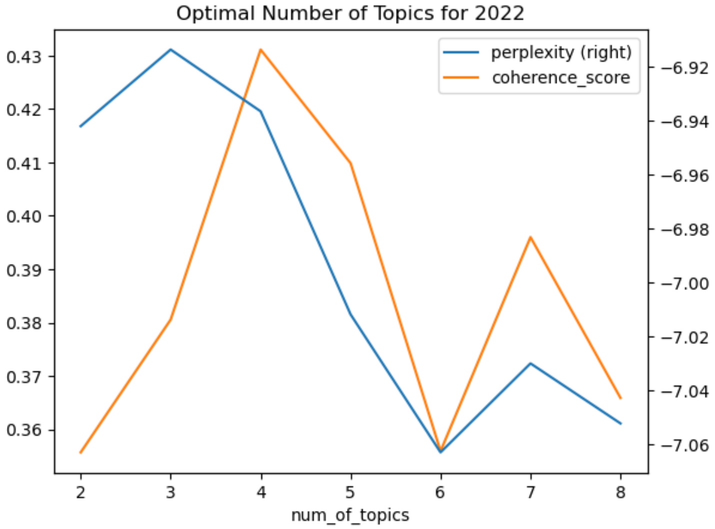


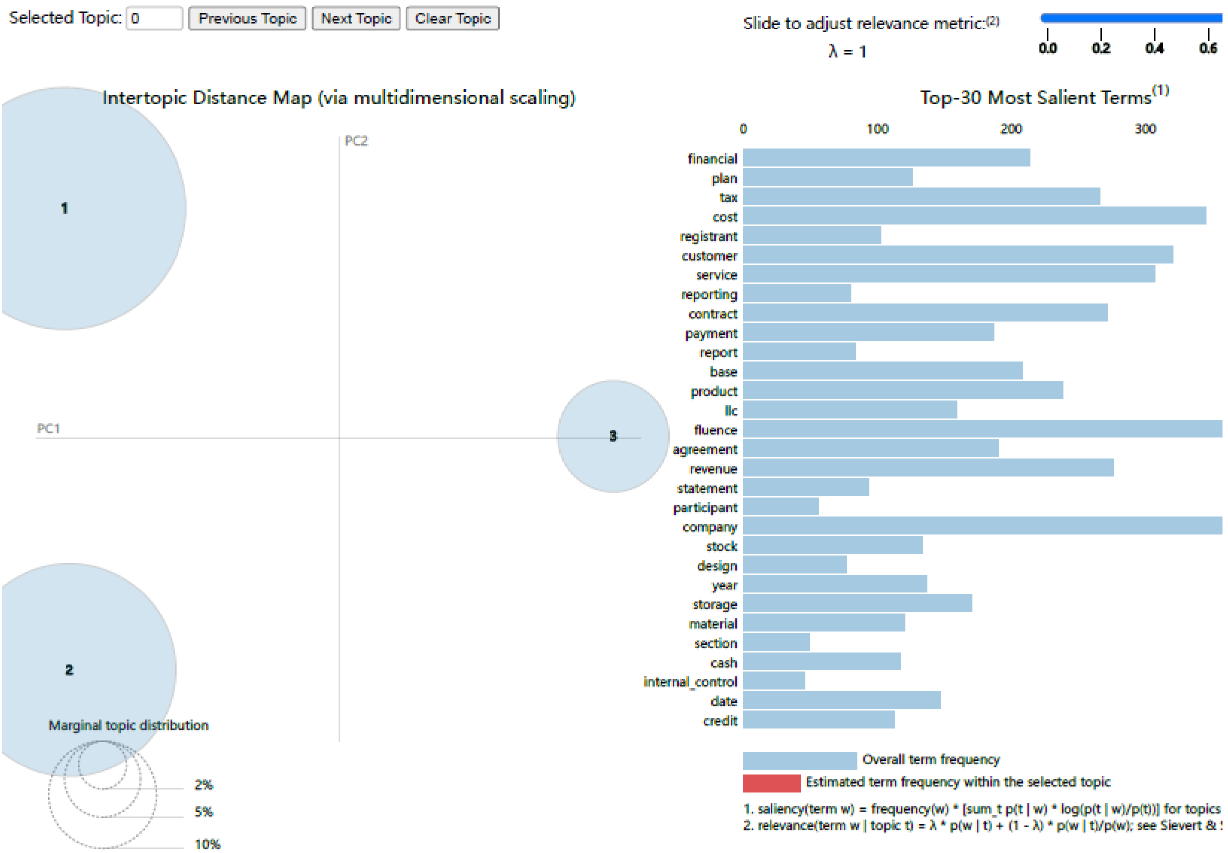
Figure 6: Number of Topics

In the figure below, we take 2022 as an example and present the LDA analysis results using an Intertopic Distance Map (via multidimensional scaling). This interactive visualization allows us to observe the relative positioning and separation of the identified topics in a two-dimensional space. The resulting map is shown as follows:

It is worth noting that while the LDA topic modeling results for 2023 and 2024 remained consistent—both identifying three main topics clustered in similar regions of the intertopic distance map—the year 2022 revealed an additional topic, making a total of four topics. This extra topic appears in the second quadrant of the PC1-PC2 coordinate system and is labeled as Topic 3, which does not appear in the reports of the following two years.

From the word frequency analysis associated with this topic, we can observe that it is primarily centered around terms such as `"contract"`, `"service"`, and `"revenue"`. This suggests that the topic may correspond to specific contractual or revenue-related developments in 2022, possibly tied to a particular business line or a set of business agreements that year.

Returning to the word frequency bar charts discussed earlier, we find that the term `"contract"` ranked 15th in 2022 but dropped to 29th in 2023 and 30th in 2024, which aligns well with the disappearance of the associated topic in the later years.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & !

Figure 7: LDA 2024

The presence of an additional topic in 2022, focused on "contract", "service", and "revenue", implies that the company experienced notable contractual activities or revenue-related changes during that year. These could be attributed to major new deals, changes in service agreements, or strategic shifts affecting the company's revenue model.

The sharp decline in the prominence of this topic in subsequent years indicates that these activities were likely temporary or non-recurring. For investors and analysts, this pattern may imply that 2022 saw a short-term boost in revenue or contractual commitments, which may not have had a sustained impact on Fluence's long-term financial performance. It is therefore advisable to further investigate whether these developments led to enduring changes in business operations or were isolated events.

4. ESG Sentiment and Keyword Analysis

Research Objective

Count the frequency and proportion of ESG keywords that appear in the 10-K each year to see which issues have risen/fallen most significantly.

To address this problem, we need to incorporate a language package or library capable of recognizing ESG-related content in reports. Therefore, the first step is to introduce an ESG-specific language model.

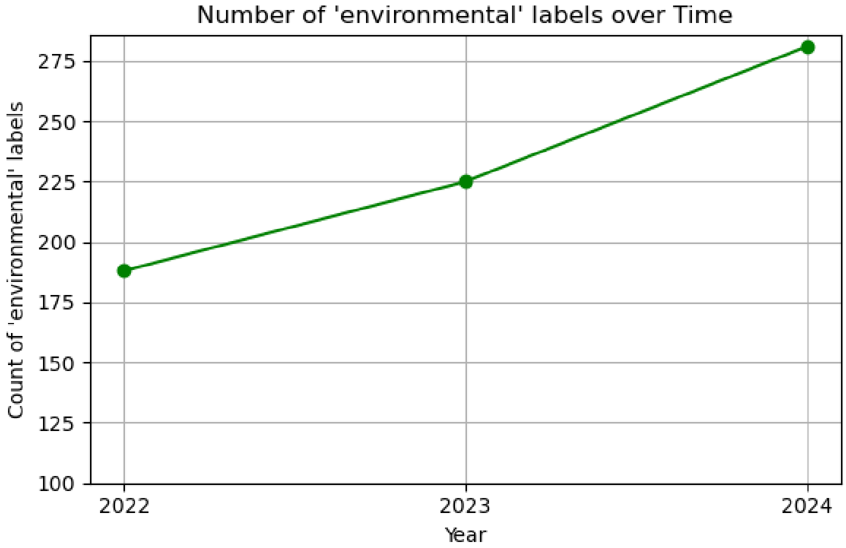
As with the earlier stages, I utilized the capabilities of Hugging Face by integrating an ESG text classification model along with a sentiment analysis model. These models help identify ESG-related sections in the reports and analyze the sentiment behind them. Below is an overview of how each model was applied and the specific problems they address:

- **Environmental Classification:** Identifies sentences related explicitly to environmental topics within ESG reports.
- **Sentiment Analysis:** Analyzes overall sentiment expressed in sentences from ESG reports, categorizing them into positive, negative, and neutral.

This integrated analysis provides a comprehensive view of ESG disclosure patterns, and the sentiment conveyed through corporate sustainability communications.

We firstly load ESG-related language models and preprocesse report sentences to identify environmental content and verify model labels before performing classification.

Then ,we loop through the same ESG report sentences to classify each sentence's sentiment (`positive`, `neutral`, `negative`). Count and display sentiment distributions for each year. Results shown in the chart below:



The rising count of sentences classified as "environmental"—from 188 (2022) to 225 (2023) and 281 (2024)—demonstrates Fluence Energy's increasing emphasis on environmental topics in ESG reporting. This growth reflects stronger internal prioritization, stakeholder pressure, and heightened sustainability awareness. Sentiment analysis further shows a shift toward more neutral and positive tones, suggesting greater confidence and a proactive stance in ESG communications. Together, these trends indicate improving ESG maturity and alignment with global sustainability expectations.