NLP Analysis of Trent (Zudio) FY24-25 Annual Report

Roll No.: L05	**Name:** k John Austin
Program : B.tech CSEDS	**Sem/Year :** VII / 4th Year

Abstract

This report presents an end-to-end Natural Language Processing (NLP) analysis of the Trent Limited (Zudio) FY24-25 annual report. Using document parsing, text cleaning, sentiment analysis, keyword and frequency analysis, and topic modeling, we extract insights about strategic themes, financial emphasis, and communication tone. Results show recurring themes around company performance, governance, financial statements, and growth, consistent with typical annual report narratives.

Keywords

NLP, Annual Report, Sentiment Analysis, Topic Modeling, WordCloud, TF-IDF, LDA, Trent, Zudio

1. Introduction

Corporate annual reports are rich narrative documents that combine regulatory disclosures with management discussion, strategy, and performance commentary. Applying NLP to such documents provides a scalable way to summarize themes, identify sentiment, and detect important keywords beyond manual reading.

This project analyzes trent_fy24-25_annual_report.pdf using a reproducible notebook (NLP_Mini_Project.ipynb). We mirror common academic mini-project structure: problem framing, methodology, implementation, results, discussion, and conclusions.

Objectives

- Extract and clean text from the Trent (Zudio) FY24-25 annual report.
- Quantify sentiment at sentence and document levels.
- Identify frequent words, salient keywords, and visualize via WordCloud and bar charts.
- Model underlying themes with LDA topic modeling.

2. Dataset and Resources

- Primary document: trent_fy24-25_annual_report.pdf
- Environment: Python, Jupyter Notebook (NLP_Mini_Project.ipynb).
- Libraries: pdfplumber, nltk, textblob, gensim, scikit-learn, wordcloud, matplotlib, seaborn, pyLDAvis, pandas, spacy (installed; not central to final pipeline).

3. Methodology

Document Ingestion

- Load PDF with pdfplumber and extract text page by page into a pandas DataFrame with columns: page_num, text.

Preprocessing

- Lowercasing, removal of digits and punctuation, whitespace normalization.
- Tokenization and stopword removal (nltk.stopwords).
- Result stored as clean_text per page.

Sentiment Analysis

- Sentence tokenization (sent_tokenize).
- Per-sentence polarity and subjectivity using TextBlob.
- Aggregate metrics: average polarity and counts of positive sentences.

Lexical Statistics and Visualization

- Word frequency via collections.Counter/pandas.value_counts.
- Visualization: bar charts of top-N words and a WordCloud of the full corpus.

Vectorization and Topic Modeling

- Feature extraction: TfidfVectorizer and CountVectorizer (max features ~500–2000 as used in the notebook).
- Topic modeling: gensim LDA with adjustable num_topics and training passes.
- Topic visualization and interpretability: pyLDAvis in the notebook.

4. System Design and Implementation

Implementation via Notebook (NLP_Mini_Project.ipynb)

1. Install/import dependencies and download NLTK corpora.
2. Extract pages from the PDF with pdfplumber.
3. Create DataFrame; derive clean_text using regex, tokenization, and stopword removal.
4. Sentence-level sentiment extraction and aggregation.
5. Word frequency and WordCloud visualization.
6. TF-IDF and Count vectorization to build document-term matrices.
7. gensim LDA model training (e.g., num_topics=10, passes=10).
8. pyLDAvis interactive topic exploration.
9. Export artifacts: trent_cleaned_text.csv and trent_sentence_sentiments.csv.

Key Implementation Notes

- Robust NLTK downloads: handle punkt_tab vs punkt gracefully.

- Cleaning choices (digits/punctuations removal) favor topic clarity; numeric analysis is out of scope.
- Single-document LDA is less stable than corpus-level modeling; interpret multiple runs and tune topic numbers.

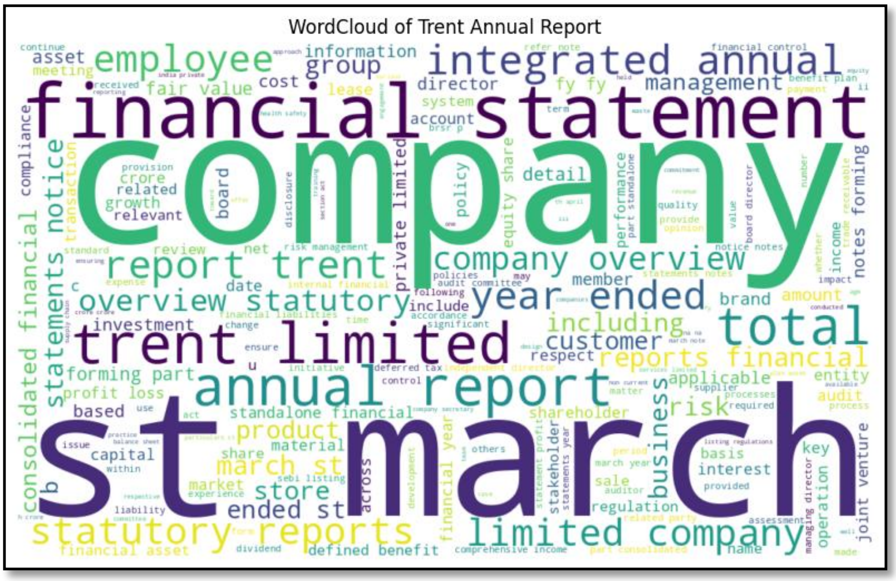
5. Results

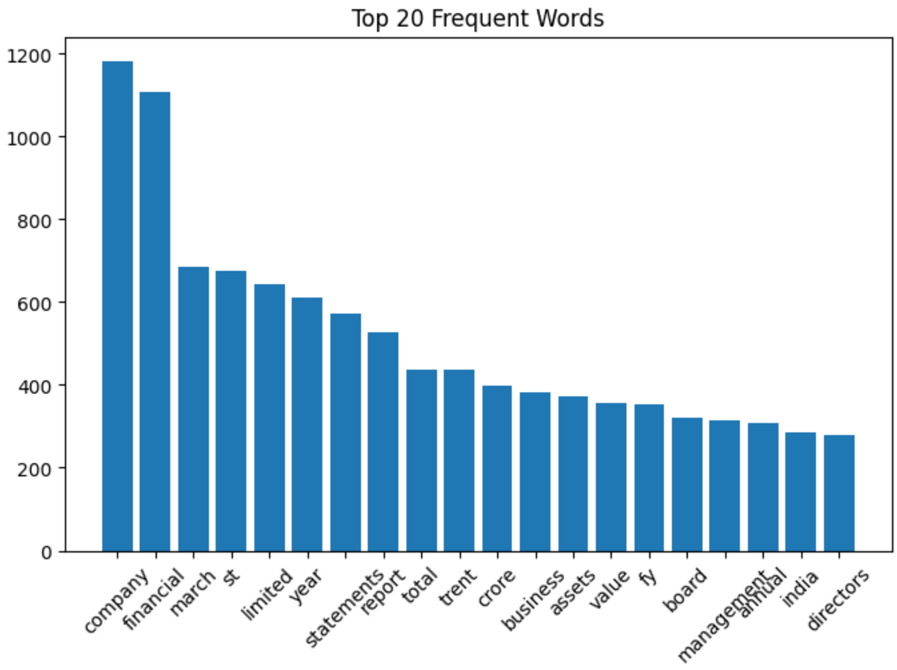
Descriptive Metrics

- Example metrics observed on the report include total pages, word and sentence counts, average polarity, and unique words. These provide a quick orientation to document size and tone.

Word Frequency and WordCloud

- Frequent tokens highlight governance, financial reporting, and company references, reflecting typical annual report structure.

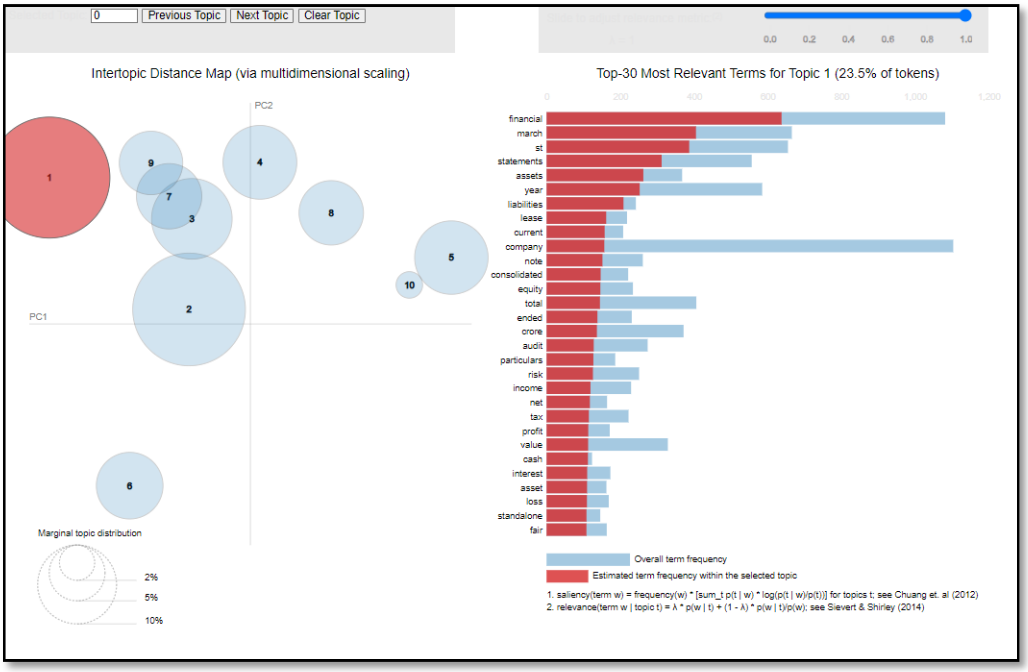


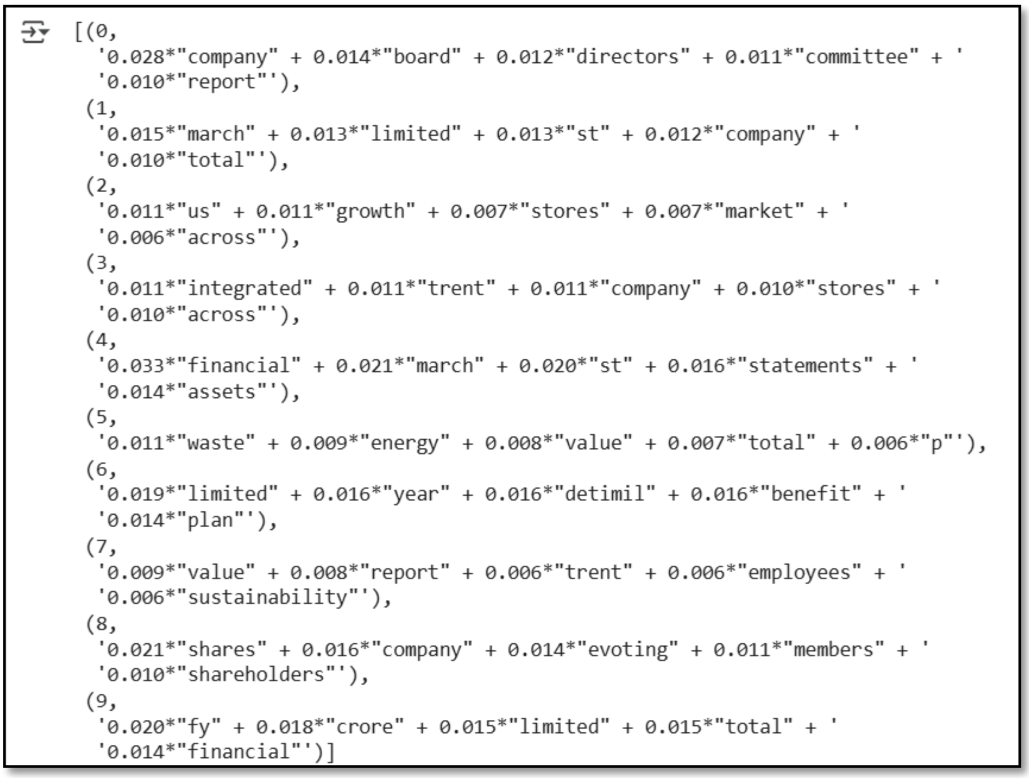


Topics (Notebook gensim LDA Snapshot)

The notebook surfaced interpretable topics indicative of sections such as governance, financial statements, and growth (illustrative top words shown during runs):

- Topic examples included: "company", "board", "directors", "committee", "report"; financial cluster around "financial", "statements", "assets"; governance/AGM cluster around "shares", "evoting", "members", "shareholders"; growth/market cluster around "growth", "stores", "market", "across".


0
Previous Topic
Next Topic
Clear Topic
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (23.5% of tokens)
0
200
400
600
800
1,000
1,200
financial
march
st
statements
assets
year
liabilities
lease
current
company
note
consolidated
equity
total
ended
crore
audit
particulars
risk
income
net
tax
profit
value
cash
interest
asset
loss
standalone
fair
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



```
[(0,
  '0.028*"company" + 0.014*"board" + 0.012*"directors" + 0.011*"committee" + '
  '0.010*"report"'),
 (1,
  '0.015*"march" + 0.013*"limited" + 0.013*"st" + 0.012*"company" + '
  '0.010*"total"'),
 (2,
  '0.011*"us" + 0.011*"growth" + 0.007*"stores" + 0.007*"market" + '
  '0.006*"across"'),
 (3,
  '0.011*"integrated" + 0.011*"trent" + 0.011*"company" + 0.010*"stores" + '
  '0.010*"across"'),
 (4,
  '0.033*"financial" + 0.021*"march" + 0.020*"st" + 0.016*"statements" + '
  '0.014*"assets"'),
 (5,
  '0.011*"waste" + 0.009*"energy" + 0.008*"value" + 0.007*"total" + 0.006*"p"'),
 (6,
  '0.019*"limited" + 0.016*"year" + 0.016*"detimil" + 0.016*"benefit" + '
  '0.014*"plan"'),
 (7,
  '0.009*"value" + 0.008*"report" + 0.006*"trent" + 0.006*"employees" + '
  '0.006*"sustainability"'),
 (8,
  '0.021*"shares" + 0.016*"company" + 0.014*"evoting" + 0.011*"members" + '
  '0.010*"shareholders"'),
 (9,
  '0.020*"fy" + 0.018*"crore" + 0.015*"limited" + 0.015*"total" + '
  '0.014*"financial"')]
```

Sentiment

- Average polarity is near-neutral as expected for formal corporate disclosures.
- Positive sentences often appear in performance highlights; negative or lower polarity sentences in risk or cautionary statements.

Conclusion :

This mini-project demonstrates a complete notebook-driven NLP pipeline to analyze the Trent (Zudio) FY24-25 annual report. By extracting text from PDF, cleaning and tokenizing it, computing sentence-level sentiment, exploring word distributions, and modeling topics with LDA, we surface the document's dominant themes and overall tone. The topics align with expected sections such as governance, financial statements, and growth, while sentiment remains near-neutral as typical of formal disclosures. The approach is reproducible, extensible to other annual reports, and can be enriched with finance-aware preprocessing and multi-year comparisons in future work.