Topic Modeling Across S&P 500 Earnings Calls

Team 202 Mark Emmenegger | Talha Sumra | Sumit Singh | Kenneth Thorburn | Nathaniel Beare | Gordon Garisch

Motivation/ Introduction

What is the Problem?
Earnings calls from S&P 500 companies contain a wealth of unstructured financial, strategic, and operational insights. Manually extracting recurring themes from tens of thousands of transcripts is infeasible. We aim to automate this.

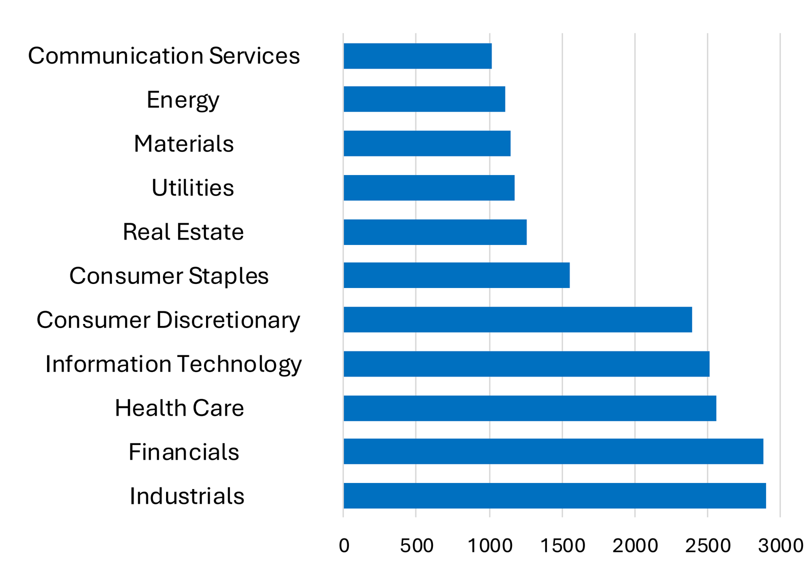
Why is it important?
These calls influence investment decisions and market sentiment. Understanding trends in key themes, such as technology, inflation, ESG, or AI, can empower investors, analysts, and policymakers to make more informed decisions. Our tool thus aims to significantly enhance users' market intelligence. It targets institutional and retail investors, market analysts, executives, and policy makers.

Data

How Was It Collected?
- Earnings call transcripts (Q1 2014–Q4 2024) from API Ninjas.
- S&P 500 company metadata from GitHub repositories.

Data Characteristics
- Size: >20,000 transcripts, ~2GB.
- Structure: Each transcript ≈ 80 speaker statements
- Temporal: Quarterly over 10 years.
- Sectors: All GICS sectors represented.

# of Earnings Calls across Sectors, 2014-2024



Our Approaches

What are the Approaches?
We evaluate three topic modeling techniques LDA (Latent Dirichlet Allocation) , NMF (Negative Matrix Factorization) and BERTopic on over 20,000 earnings call transcripts. We enrich these topics using clustering, LLM-based naming, and present them in an interactive dashboard.

How do they work?

1

Preprocessing:
- Text Cleaning: Filter unnecessary words and sentences
- Lemmatization: Convert words to their base forms
- Creating Document-Term-Matrix

Azure

2

Topic Modeling:
- LDA: Bayesian probabilistic model
- NMF: Matrix factorization method
- BERTopic: Embedding + clustering + term weighting

Azure

3

Hierarchical clustering and Naming:
- Hierarchical method (dendrogramm) to build topic hierarchy
- GPT-4 reviews hierarchical clustering and assigns interpretable labels to topics and clusters, as a human analyst would

OpenAI

4

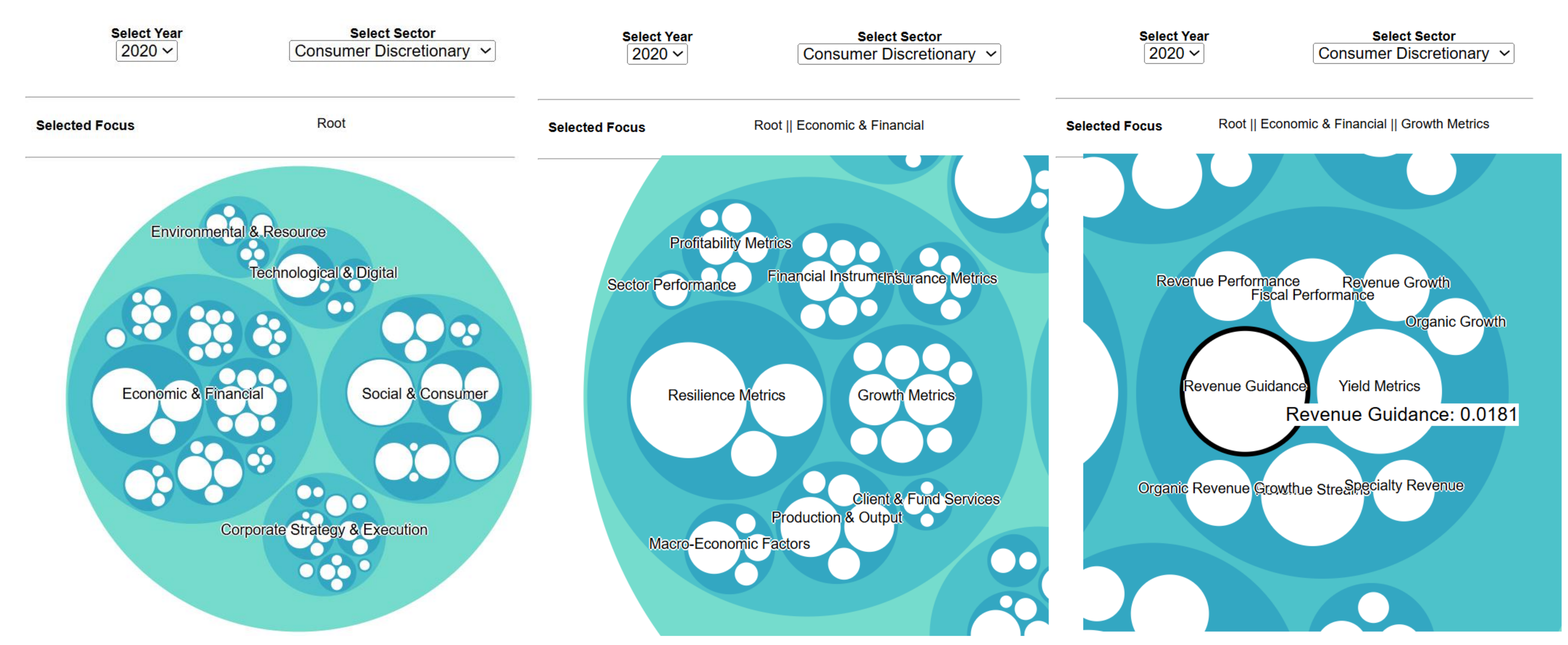
Visualization:
- Two Tableau dashboards with:
- Embedded D3.js-based nested circle packing
- Time-series plots, bar charts, top topic list

+ a b l e a u

Why they work well?
Combining topic modeling (LDA) with scalable LLM-based interpretation produces meaningful, human-readable outputs. Hierarchical clustering uncovers structure in the otherwise flat topic landscape.

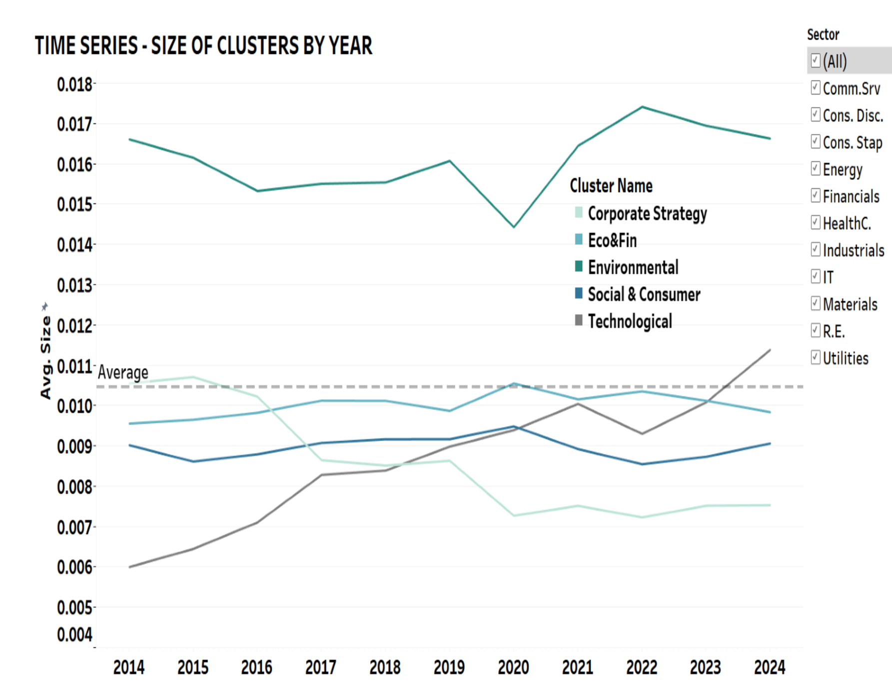
What Is New?
- LLM-guided topic clustering and naming pipeline: Combines classical hierarchical clustering algorithm with GPT-4 refinement to automatically generate interpretable, hierarchical themes at scale.
- Circle packing chart: Interactive visualization of a multi-level topic structure created in D3.js and embedded in Tableau.

Selected Dashboard Element 1: Circle packing chart (Zoom)

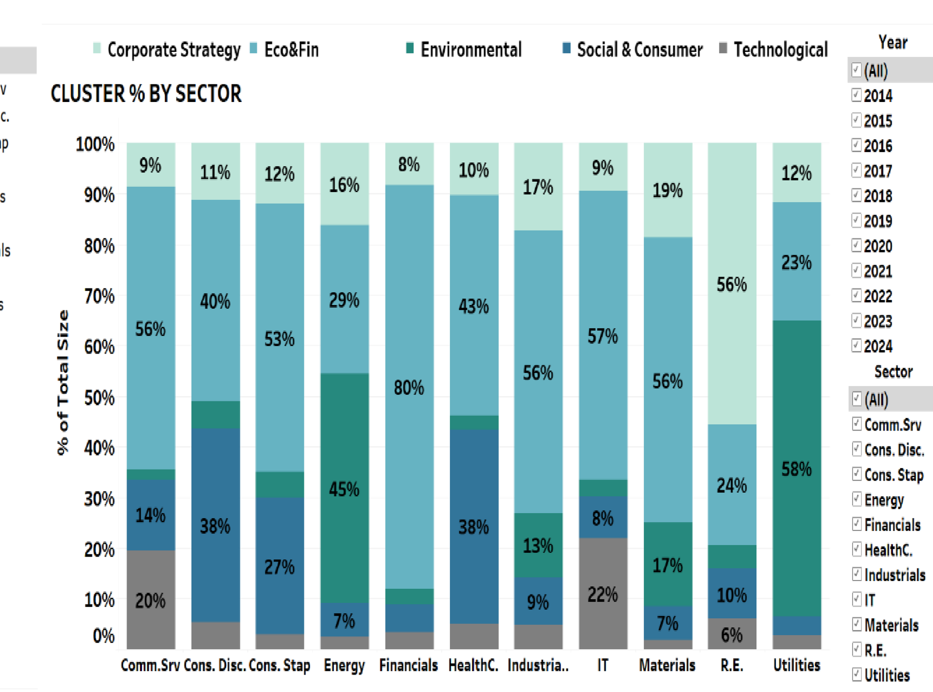


Selected Dashboard Elements 2 and 3: Time Series Chart and Stacked Bar Chart

Evolution of Themes Prominence over Time



Distribution of Themes across Sectors



Experiments and Results

Evaluation Approach
- Model quality: Coherence score, topic diversity, intruder detection.
- Survey: Topic naming, cluster interpretability, alignment with ground truth, dashboard usability and appearance.

Model quality results for our 2 Candidate Models tuned on a Range of Topics

Model			Quant. measures		Qualit. measure
Type	No. Topics	Data	umass	topic diversity	Intruder analysis
LDA	50	sample	-1.12	0.84	4
NMF	50	sample	-1.04	0.84	4
BERTopic	50	sample	-1.24	0.96	3
LDA	100	sample	-0.99	0.64	5
LDA (*)	100	full	-0.98	0.66	5
NMF	100	sample	-1.01	0.65	4
BERTopic	100	sample	-1.22	0.86	3
LDA	150	sample	-0.99	0.53	3
NMF	150	sample	-0.99	0.58	3
BERTopic	150	sample	-1.24	0.71	2
LDA	200	sample	-0.99	0.46	3
NMF	200	sample	-1.01	0.53	2
BERTopic	200	sample	-1.20	0.58	2

Notes: (*) Final model. Sample data = random sample of 2,000 earnings calls (approx. 10% of the data); used for tuning due to run time issues. Full data = full sample of 20,523 earnings calls. Intruder Analysis: Result rated from 1 (worst) to 5 (best) by the modelers.

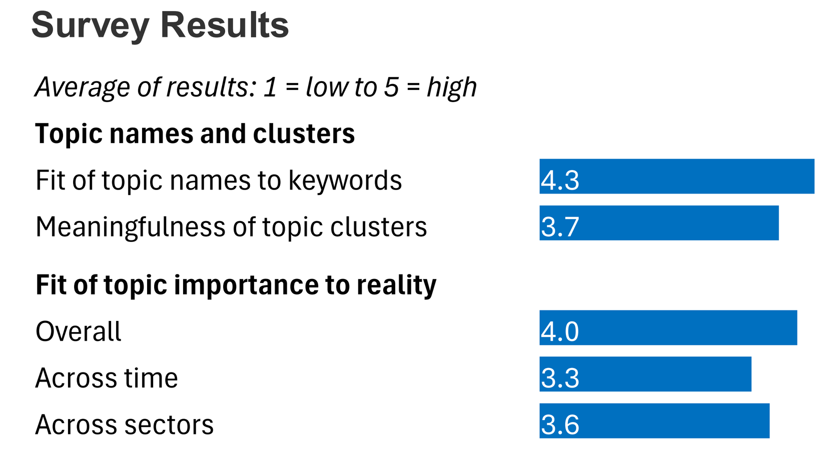
Examples from Intruder Analysis for LDA w/ 100 topics:

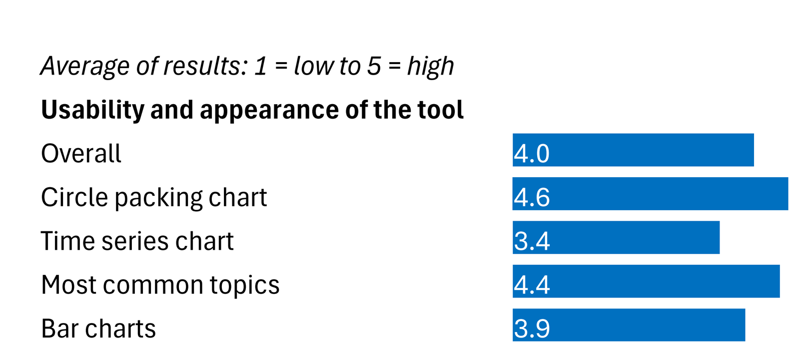
Topic 1: ai | design | software | data|innovation
Intruder: lower

Topic 2: aircraft | united | fleet | delivery|international
Intruder: sale increase

Topic 3: Marketing |user | app | mobile | advertising
Intruder: consolidated

Survey Results





Notes: Survey conducted with 7 subject matter experts from finance and economics

Results Summary
- Best Model: LDA (100 topics), highest coherence (–0.99), good diversity (0.64), best in intruder analysis, intuitive explanation
- LLM Labeling: Topic name fit rated 4.3/5.
- Clusters: Rated 3.7/5 for logical grouping.
- Tool Usability: Rated 4.0/5.
- Top Chart: Circle packing (4.4/5), lowest: Time series chart (3.4/5).

Comparison to Others
- Contrary to Twitter data studies (Egger, R., & Yu, J. 2022), LDA outperforms BERTopic and NMF for longer, structured texts.