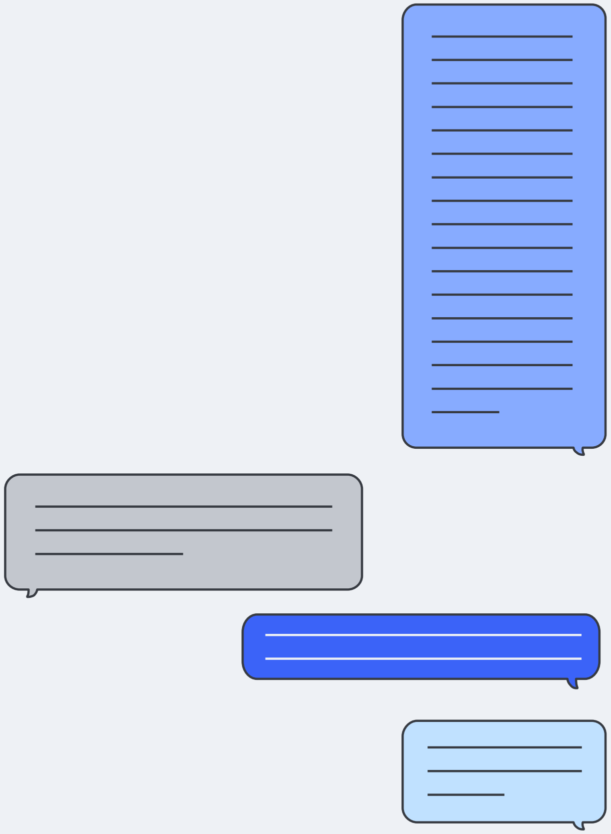
NLP in Finance

Content Extraction and Sentiment Analysis for Earning Call Transcripts

Team 6: Sophia Moore, Eric Chen, Dyuthi Dinesh, Maud Yaiche, and Michael Bartell



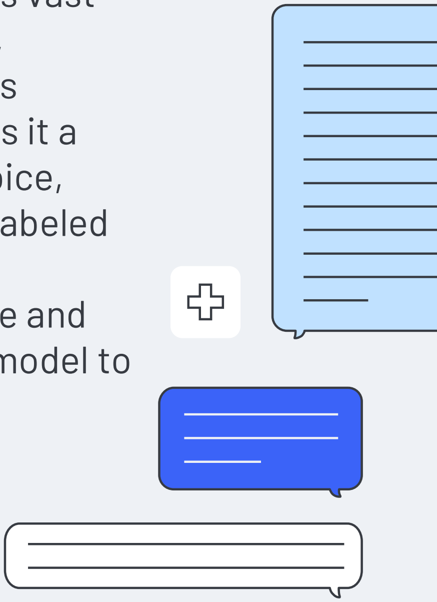
Introduction

Our objective

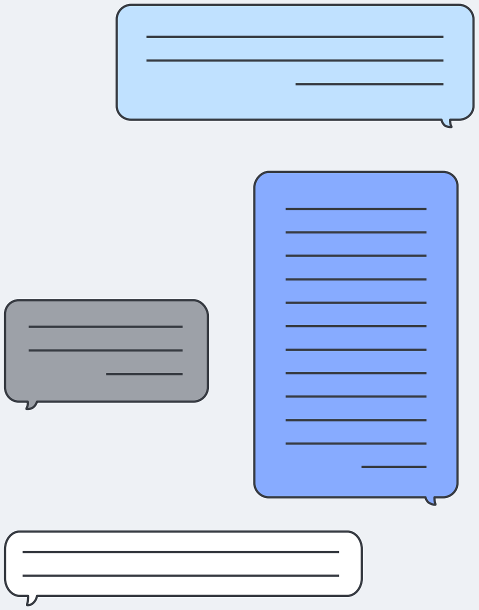
The objective of this project is to analyze corporate earnings calls to distinguish between idiosyncratic and systematic guidance to provide company-specific insights.
The goal is to provide **a succinct summary of topics** covered and evaluate the **company's sentiment toward both its own performance and its sector**.

Our idea

Unsupervised learning leverages vast amounts of unlabeled text data, focusing on uncovering patterns manual intervention. This makes it a more practical and scalable choice, especially for scenarios where labeled data is sparse or unavailable.
Our goal is to find the best viable and well-performing unsupervised model to train our data.



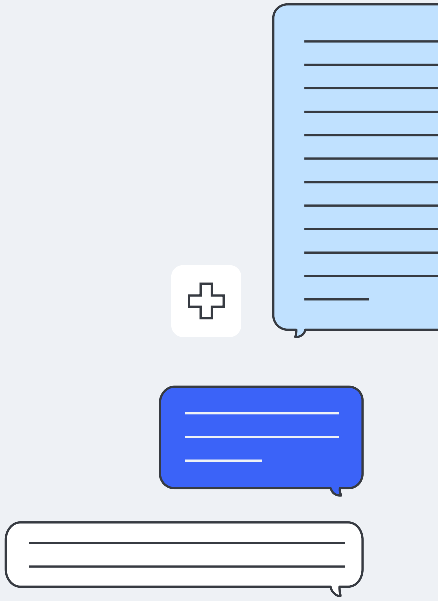
Data Collection and Initial Cleaning



Data collection

Getting the Earning calls from key healthcare companies

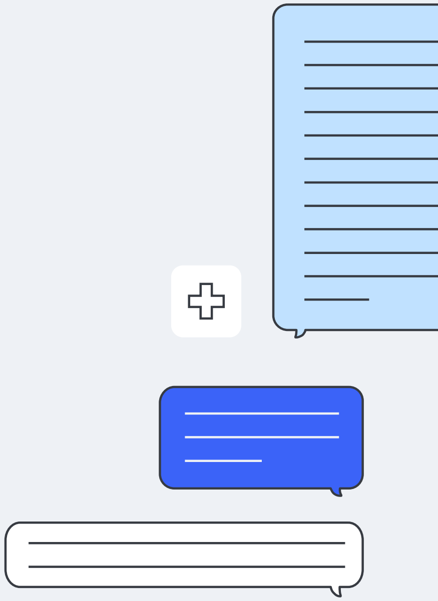
- **Wikipedia** list of top health care companies in the US.
- **Facset** for all the Transcript of the earning calls (failed webscraping, uploading by hand)



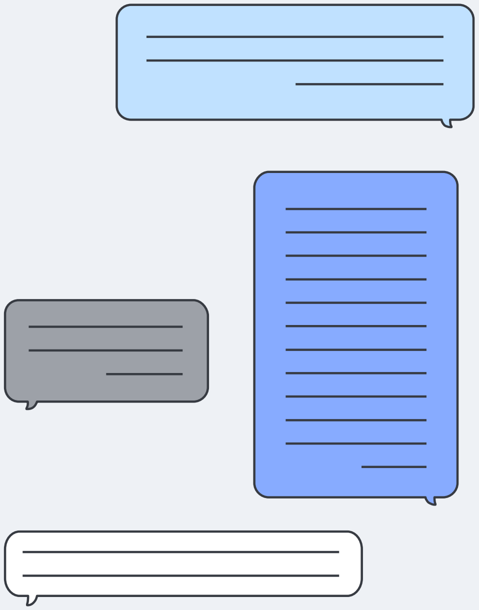
Data cleaning

Web scraped data cleaning

- Cleaned to obtain text files from a random set of **10 HealthCare companies** for the time period of **Jun-Aug 2024**.
- The text files were first **tokenized** into *sentences*.
- Each sentence was cleaned using **regular expressions** functions, and **stopword removal**.
- Text Lemmenization was conducted on the cleaned sentences (for context extraction).



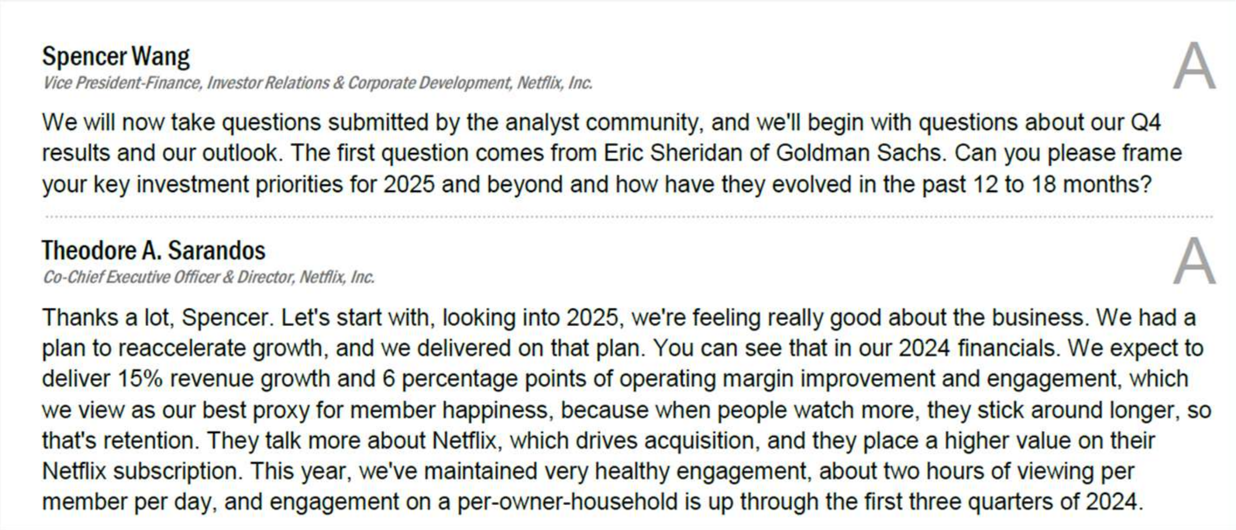
Content Extraction



Earning Call Transcript Ex.

- A typical earning call transcript has a mix of actual content rich material and unimportant transitional or dialogue like sentences.

- Depending on the web script source, the format for such transcripts are not consistent which added to the challenge of data cleaning.



Spencer Wang
Vice President-Finance, Investor Relations & Corporate Development, Netflix, Inc.

We will now take questions submitted by the analyst community, and we'll begin with questions about our Q4 results and our outlook. The first question comes from Eric Sheridan of Goldman Sachs. Can you please frame your key investment priorities for 2025 and beyond and how have they evolved in the past 12 to 18 months?

Theodore A. Sarandos
Co-Chief Executive Officer & Director, Netflix, Inc.

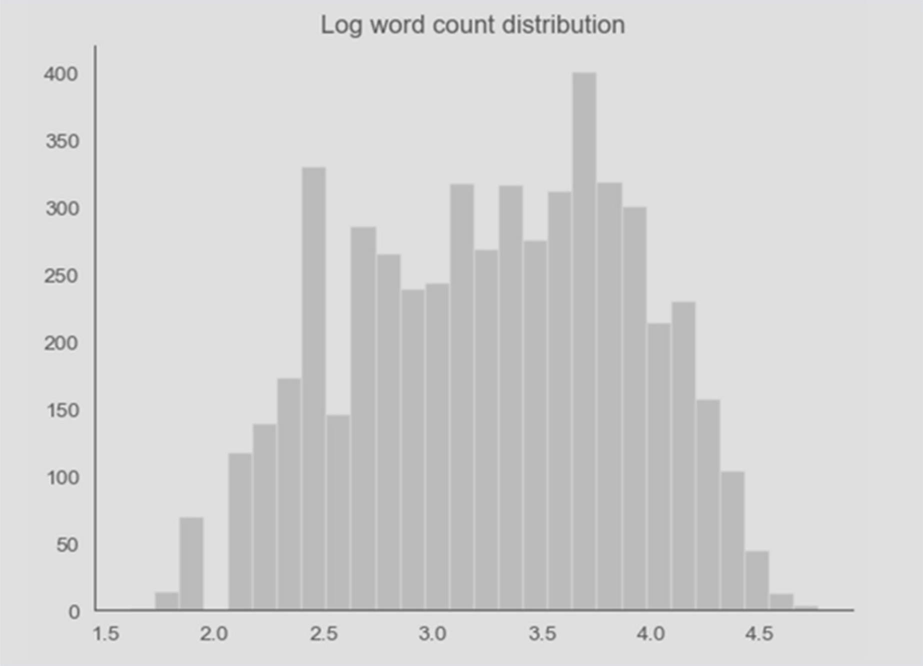
Thanks a lot, Spencer. Let's start with, looking into 2025, we're feeling really good about the business. We had a plan to reaccelerate growth, and we delivered on that plan. You can see that in our 2024 financials. We expect to deliver 15% revenue growth and 6 percentage points of operating margin improvement and engagement, which we view as our best proxy for member happiness, because when people watch more, they stick around longer, so that's retention. They talk more about Netflix, which drives acquisition, and they place a higher value on their Netflix subscription. This year, we've maintained very healthy engagement, about two hours of viewing per member per day, and engagement on a per-owner-household is up through the first three quarters of 2024.

Data Cleaning and Exploration

- Combine all the corpuses for a company in one text and dissect it by paragraph.

- Use regex to remove odds symbols coming out of data scraping, remove stop words, and apply lemma to further shrink possible token.

- Remove any paragraphs fewer than 25 characters.

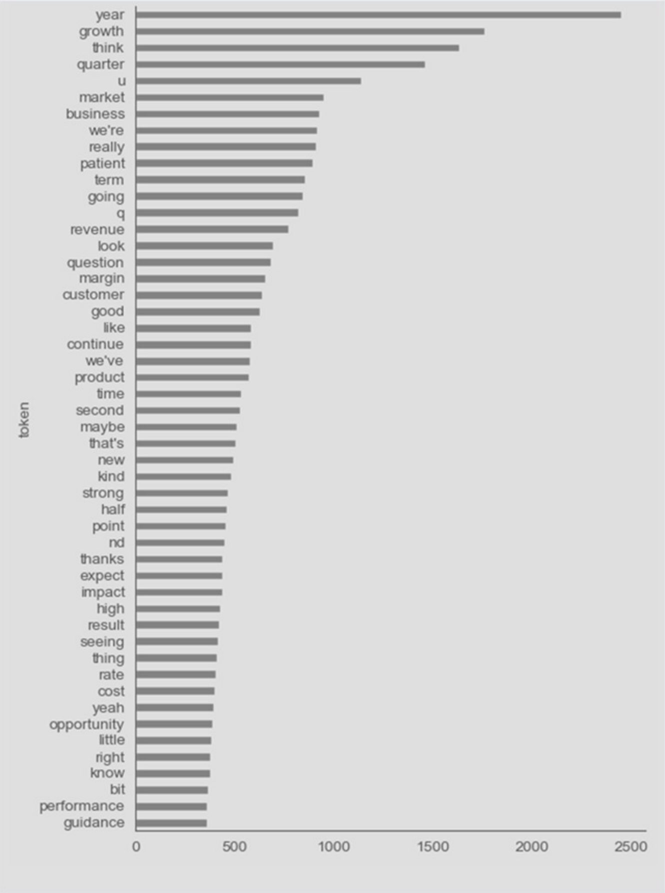


Corpus Statistics
Count - 5039
Mean - 31.78
Standard dev. - 19.22

Data Exploration cont.

We analyzed single-token frequency, highlighting the 50 most popular tokens. Common words like *growth*, *market*, and *revenue* dominate.

The frequency distribution shows that overused words, particularly the top five, contribute little to context from a TF-IDF perspective, while stabilization around the midpoint (~25) informs the next step of data vectorization.



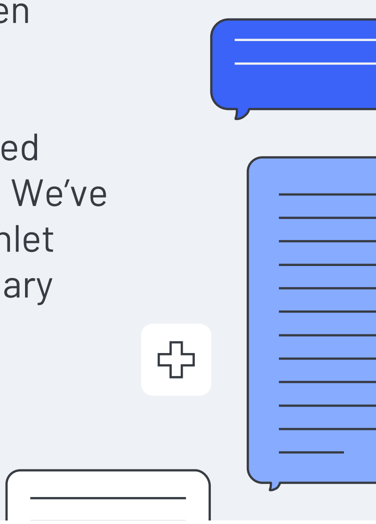
Supervised/Unsupervised Learning

Supervised Learning

- Relies heavily on large, human-labeled datasets, which are both expensive and time-intensive to create.
- Tasks such as sentiment analysis, named entity recognition, and translation require meticulous annotation, making scalability a challenge.

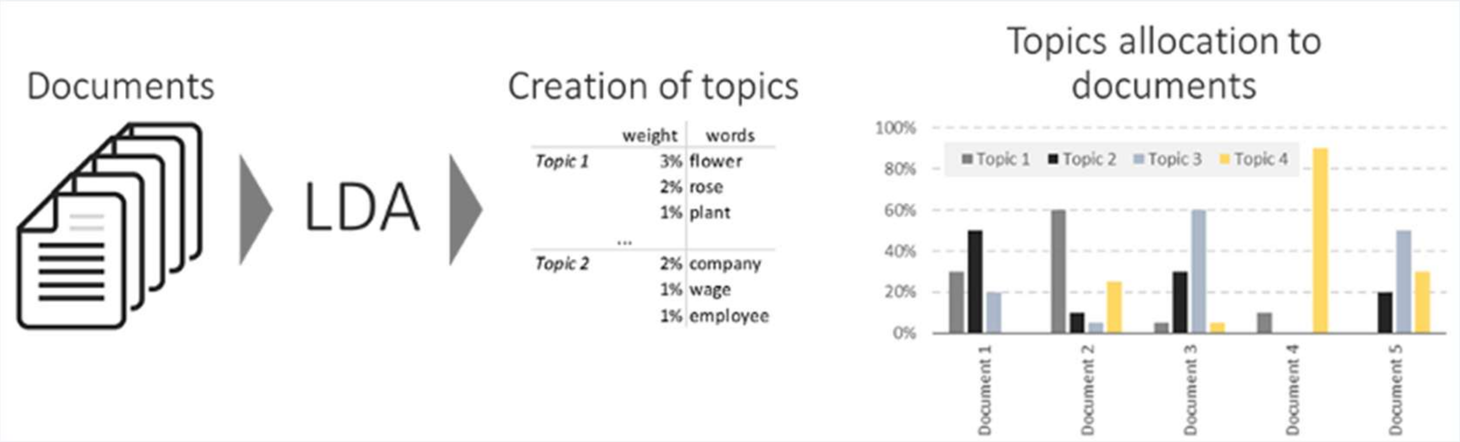
Unsupervised Learning

- Works with vast amounts of unlabeled text data, uncovering patterns and latent structures without manual intervention.
- This approach is more practical and scalable, especially when labeled data is sparse or unavailable.
- For our project, unsupervised models are a logical choice. We've decided to use Latent Dirichlet Allocation (LDA) as our primary model.



Latent Dirichlet Allocation (LDA)

- LDA is a hierarchical Bayesian model that assumes topics are probability distributions over words, and documents are distributions over topics.
- More specifically, the model assumes that topics follow a sparse Dirichlet distribution, which implies that **documents cover only a small set of topics, and topics use only a small set of words frequently.**
- Its unsupervised nature means there are no clear metrics like accuracy or confusion matrices to evaluate its results. This makes outcomes harder to interpret, though human evaluations of topics are often impractical and costly.

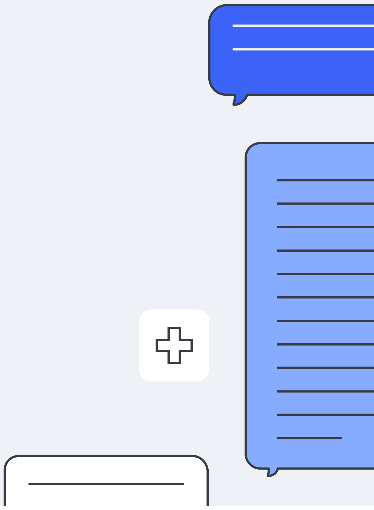


Content Extraction

- Create a document-term matrix with terms appearing in 0.5–50% of documents, reducing the token count from over 5,000 to around 980.
- For parameter tuning, we focused on two key settings: the number of topics and the number of passes.

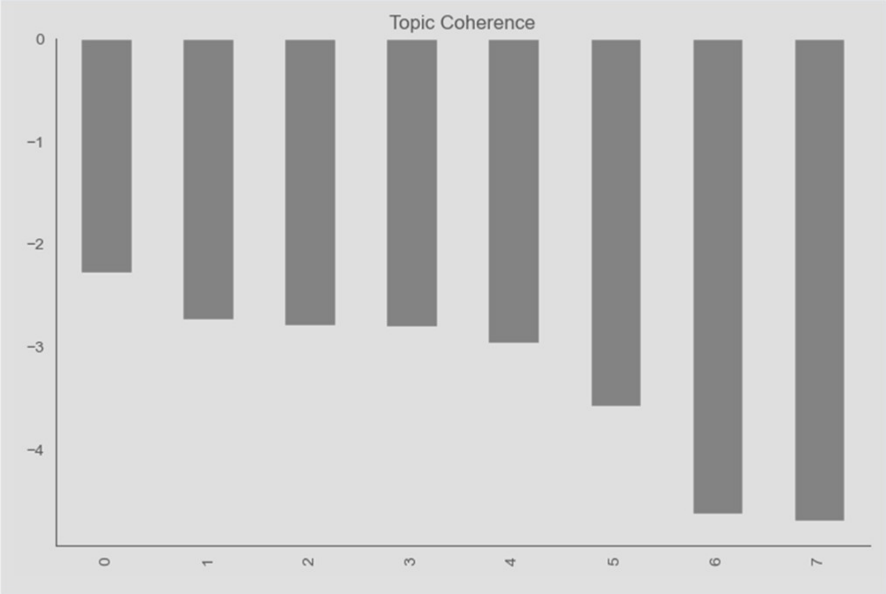
Hyperparameter tuning is more complex in unsupervised models due to the lack of universal metrics. For our LDA model, we used the coherence score as the primary evaluation metric. This score measures how semantically related the words in each topic are by analyzing their co-occurrence in the original corpus. Specifically, we used the UMass coherence score, which calculates:

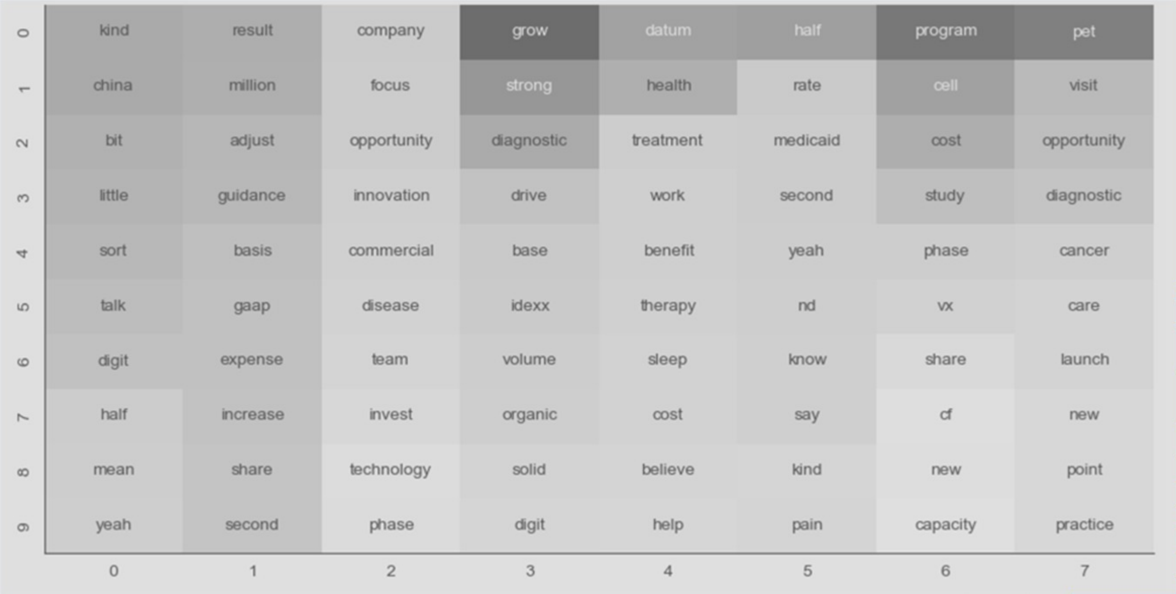
1. **Word co-occurrence** that looks at pairs of words within a topic and checks how often these words appear together in the same document in the corpus.
2. **Log-conditional probabilities**: For each pair of words, it computes the conditional probability that one word appears in a document given the presence of the other word, using logarithms to avoid scaling issues.



Content Extraction Evaluation

The gridsearch shows relatively strong coherence when we limit it to 8 topics with 25 passes given our sample size.



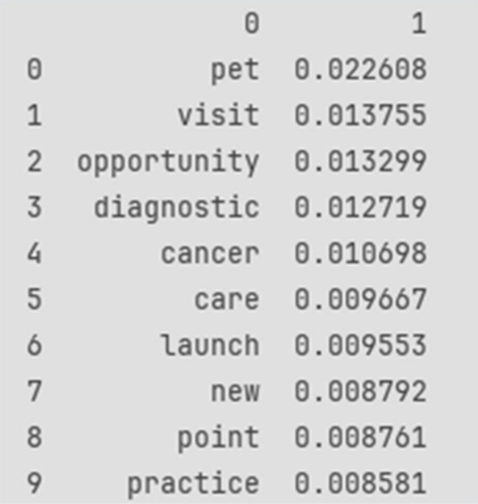


	0	1	2	3	4	5	6	7
0	kind	result	company	grow	datum	half	program	pet
1	china	million	focus	strong	health	rate	cell	visit
2	bit	adjust	opportunity	diagnostic	treatment	medicaid	cost	opportunity
3	little	guidance	innovation	drive	work	second	study	diagnostic
4	sort	basis	commercial	base	benefit	yeah	phase	cancer
5	talk	gaap	disease	idexx	therapy	nd	vx	care
6	digit	expense	team	volume	sleep	know	share	launch
7	half	increase	invest	organic	cost	say	cf	new
8	mean	share	technology	solid	believe	kind	new	point
9	yeah	second	phase	digit	help	pain	capacity	practice

Content Extraction Result Ex. I

fter a increase between and last year the net us pet population increased at the historic pre pandemic level of these pets will continue to fuel long term growth with healthcare needs throughout their lifetimes as illustrated in this graph the bars show the total clinical visits by age groups the green line is the diagnostic utilization during those clinical visits the clinical visits that include diagnostics as pets age diagnostic utilization increases the huge step up in pet population during the pandemic resulted in a flurry of clinical visits but mostly for puppies and kittens at a much lower diagnostic intensity this large cohort of pets is moving from left to right of this graph as they continue to get older while this will be an ongoing phenomena what you see here is a snapshot in time for illustrative purposes by we expect that there will be a increase in senior pet visits and a increase in geriatric pets compared to this increased diagnostic utilization as pets age amplified by the bolus of pandemic pets getting older points to a significant tailwind for our business

fter increase last year net u pet population increased historic pre pandemic level pet continue fuel long term growth healthcare need throughout lifetime illustrated graph bar show total clinical visit age group green line diagnostic utilization clinical visit clinical visit include diagnostics pet age diagnostic utilization increase huge step pet population pandemic resulted flurry clinical visit mostly puppy kitten much lower diagnostic intensity large cohort pet moving left right graph continue get older ongoing phenomenon see snapshot time illustrative purpose expect increase senior pet visit increase geriatric pet compared increased diagnostic utilization pet age amplified bolus pandemic pet getting older point significant tailwind business

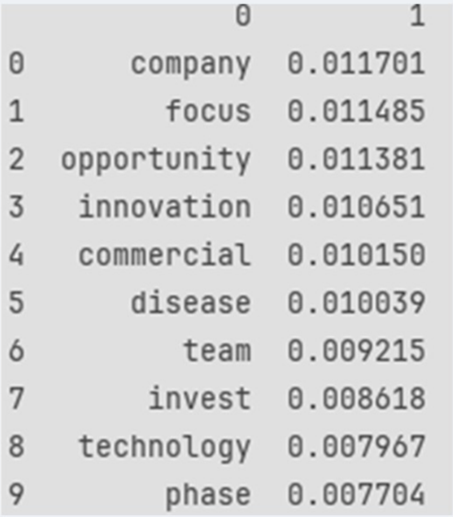


```
             0         1
0          pet  0.022608
1        visit  0.013755
2  opportunity  0.013299
3   diagnostic  0.012719
4       cancer  0.010698
5         care  0.009667
6       launch  0.009553
7          new  0.008792
8        point  0.008761
9     practice  0.008581
```

Content Extraction Result Ex. II

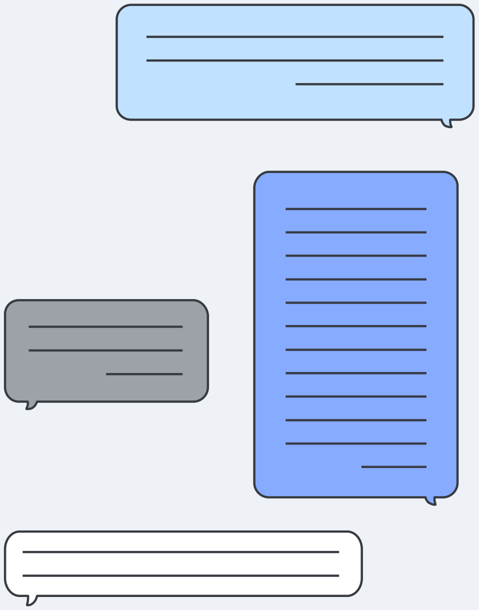
let me now give you an update on capital deployment we continue to successfully execute our disciplined capital deployment strategy which is a combination of strategic m a and returning capital to our shareholders shortly after the quarter ended we completed our acquisition of olink and it was great to welcome our new colleagues to the company earlier this month as you know olink is a leading provider of next generation proteomic solutions the addition of olink\'s proven and transformative technology is highly complementary to our industry leading mass spectrometers olink further advances our leadership as it is a great addition to our differentiated protein research ecosystem our world class commercial engine will enable us to bring this technology to scientists around the world

flet update capital deployment continue successfully execute discipline capital deployment strategy combination strategic return capital shareholder shortly quarter end complete acquisition olink great welcome new colleague company early month know olink lead provider generation proteomic solution addition olink prove transformative technology highly complementary industry lead mass spectrometer olink advance leadership great addition differentiate protein research ecosystem world class commercial engine enable u bring technology scientist world



```
             0         1
0      company  0.011701
1        focus  0.011485
2  opportunity  0.011381
3   innovation  0.010651
4   commercial  0.010150
5      disease  0.010039
6         team  0.009215
7       invest  0.008618
8   technology  0.007967
9        phase  0.007704
```

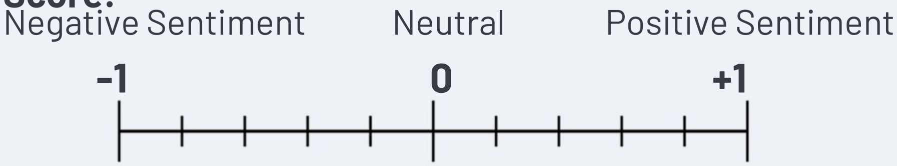
Sentiment Analysis



Sentiment Analysis

- NLTK's Vader and SentimentIntensityAnalyzer were used to get sentiment scores for each sentence in each file.
- VADER:
    - VADER first breaks down the text into individual words.
    - Then, it assigns a score to each word to identify if it is positive or negative.
    - Based on these set scores, VADER finally calculates the overall sentiment score of the text.
    - VADER also searches for modifiers that could change the meaning of neighbor words

Vader Compound Sentiment Score:



Negative Sentiment

Neutral

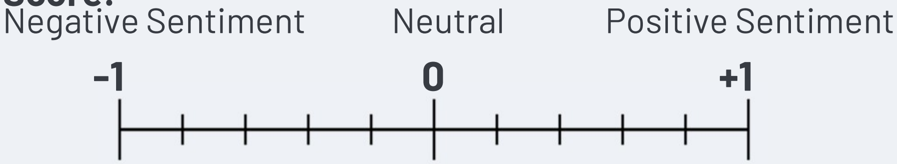
Positive Sentiment

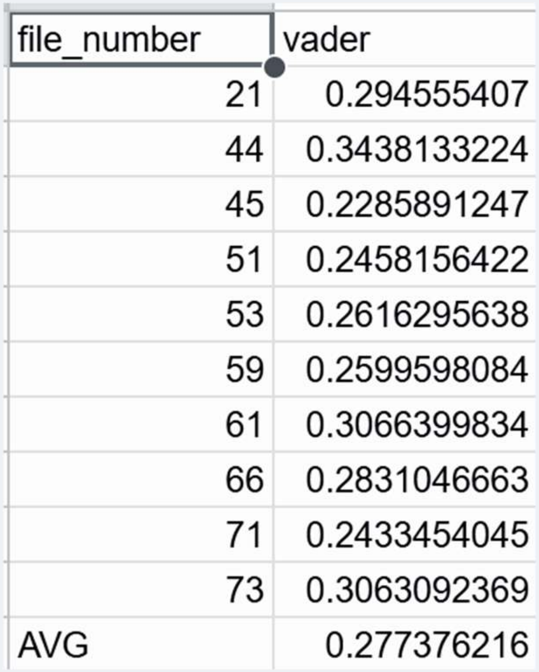
Sentiment Analysis by sentence/file

- The overall vader sentiment score for each text file was calculated by averaging the sentiment scores for **each sentence** in that text file.

- Finally, the average vader score **across all text documents** was calculated to understand the current sentiment in the healthcare sector using the earning call transcripts.

Vader Compound Sentiment Score:





file_number	vader
21	0.294555407
44	0.3438133224
45	0.2285891247
51	0.2458156422
53	0.2616295638
59	0.2599598084
61	0.3066399834
66	0.2831046663
71	0.2433454045
73	0.3063092369
AVG	0.277376216

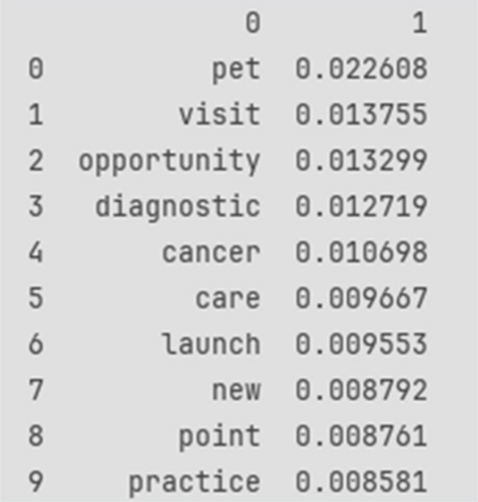
Sentiment Analysis by Topic and Word Tokenization

	Vader Score	Sentiment
Topic 0	0.9349	'neg': 0.031, 'neu': 0.779, 'pos': 0.19, 'compound': 0.9349
Topic 1	0.9493	'neg': 0.0, 'neu': 0.782, 'pos': 0.218, 'compound': 0.9493
Topic 2	0.9682	'neg': 0.063, 'neu': 0.665, 'pos': 0.272, 'compound': 0.9682
Topic 3	0.9652	'neg': 0.019, 'neu': 0.618, 'pos': 0.363, 'compound': 0.9652
Topic 4	0.8316	'neg': 0.084, 'neu': 0.716, 'pos': 0.2, 'compound': 0.8316
Topic 5	0.9393	'neg': 0.0, 'neu': 0.7, 'pos': 0.3, 'compound': 0.9393
Topic 6	0.8519	'neg': 0.05, 'neu': 0.789, 'pos': 0.161, 'compound': 0.8519
Topic 7	0.9884	'neg': 0.0, 'neu': 0.631, 'pos': 0.369, 'compound': 0.9884

Example: Combining Content Extraction and Sentiment Analysis

after a increase between and last year the net us pet population increased at the historic pre pandemic level of these pets will continue to fuel long term growth with healthcare needs throughout their lifetimes as illustrated in this graph the bars show the total clinical visits by age groups the green line is the diagnostic utilization during those clinical visits the clinical visits that include diagnostics as pets age diagnostic utilization increases the huge step up in pet population during the pandemic resulted in a flurry of clinical visits but mostly for puppies and kittens at a much lower diagnostic intensity this large cohort of pets is moving from left to right of this graph as they continue to get older while this will be an ongoing phenomena what you see here is a snapshot in time for illustrative purposes by we expect that there will be a increase in senior pet visits and a increase in geriatric pets compared to this increased diagnostic utilization as pets age amplified by the bolus of pandemic pets getting older points to a significant tailwind for our business

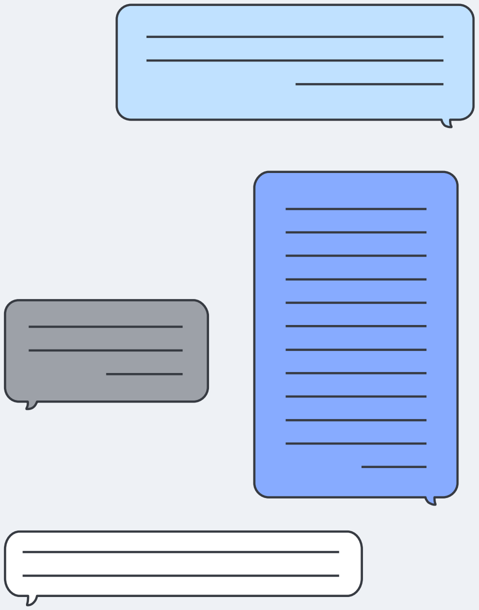
Example Lemminized Text
(on tokenized words)



```
             0         1
0          pet  0.022608
1        visit  0.013755
2  opportunity  0.013299
3   diagnostic  0.012719
4       cancer  0.010698
5         care  0.009667
6       launch  0.009553
7          new  0.008792
8        point  0.008761
9     practice  0.008581
```

Summary Sheet
Topic: Pet care visit, pet cancer related.
Sentiment: Vader score: 0.9349
'neg': 0.031, 'neu': 0.779, 'pos': 0.19, 'compound': 0.9349

Conclusion



Limitation and Future research

Data Source and Cleaning

Limitation: For the project purpose we wanted to facilitate computation time, so we only selected a limited amount of raw data

Future research:
Utilize cloud computing.
Add more earning calls and include historical transcripts to identify trends.
Winsorize by industry.

Supervised Learning

Limitation: Due to the time and resources constraints, we had to use unsupervised learning, which, though yielded satisfying results, it is not as intuitive and precise as hoped.

Future addition:
Add subject matter experts' input to tag the topics and utilize supervised learning

Market Applications

Once we incorporate more data for the model training, we want to summarize by industry and topic then compare sentiment from each company to identify any sector-level commonalities and company-level idiosyncratic outcomes. Then we will back-test against the market actual reaction as a verification of our analysis.

Current Models Tackling Finance Prediction Using NLP

ECC Analyzer

Purpose: Predict stock volatility using unstructured data from Earning Conference Calls (ECCs).

Key Features:

Utilizes Large Language Models (LLMs) for hierarchical information extraction:

1. Paragraph-Level Summaries: Extract general information
2. Fine-Grained Insights: Retrieve detailed sentences using Retrieval-Augmented Generation (RAG)

Performance:

Achieves a 27.7% reduction in Mean Squared Error (MSE) for short-term volatility predictions compared to the state-of-the-art (SOTA) models excels in Short-term forecasts (3 and 7 days)

FinBERT

Purpose: Sentiment analysis tailored to financial texts.

Unique Approach:

Pre-trained on a financial corpus (4.9 billion tokens) including earnings call transcripts, corporate reports, and analyst reports.

Incorporates a specialized financial vocabulary (FinVocab) for improved accuracy.

Performance:

Outperforms standard BERT models in financial sentiment tasks (e.g., PhraseBank, FiQA).

Combines diverse datasets for enhanced language modeling and interpretation

StockGNN

Purpose: Model relationships and dependencies in financial data.

Key Features:

Graph Neural Network (GNN):

Captures interconnections between companies and sectors.

Semantic Representations: Uses Doc2Vec embeddings to derive contextual insights from ECC transcripts.

Low Dimensional Embedding Layer:

Filters noise while retaining meaningful patterns.

Performance:

Outperforms baseline models in accuracy, precision, and recall.

Robust across multiple sectors (Finance, Health, Tech).

Demonstrates up to 23% higher recall compared to random baselines.

Thank you!