The Language of Markets: Sentiment Analysis and Key Term Extraction in Top Dow Jones Annual Reports

Presented by: Oluwatobi Ogunronbi

Introduction and Overview

- Purpose: To decipher strategic trends and sentiments within annual reports of Dow Jones Industrial Average (DJIA) listed companies.
- Scope: Analysis of textual data over the past five fiscal years, utilizing advanced computational techniques.
- Significance: Aiding stakeholders in understanding the undercurrents of corporate communication and industry-specific narratives.

Methodology Overview

In this study, we harnessed text mining techniques and sentiment analysis to scrutinize annual reports from Dow Jones Industrial Average listed companies, extracting strategic patterns and emotional tenors embedded within. The methodology integrated R's text processing packages for data preprocessing, Term Frequency-Inverse Document Frequency (TF-IDF) for emphasizing significant terms, and N-gram models to detect prevalent thematic expressions. AFINN-based sentiment analysis quantified the reports' emotional tone, with visualizations crafted using ggplot2 to illustrate trends and discrepancies across industries. This structured approach distilled voluminous textual data into clear, analyzable insights, providing a novel lens through which to view corporate strategy and

Research Questions

Research Questions (and sub-questions)	Analytical Techniques	Required Data Wrangling	Data Visualization Approaches	Notes and Special Considerations
1. What are the most frequently mentioned terms across the annual reports, and what do these suggest about overarching corporate focuses or sector trends?	Term Frequency (TF), TF-IDF, N-gram Analysis	Text extraction and cleaning, tokenization, removal of industry-common stopwords	Bar charts, frequency distributions	Incorporate n-gram analysis to capture context and phraseology indicative of industry focuses
2. What overall sentiment is conveyed in the annual reports, and how does this sentiment vary from year to year?	Sentiment Analysis	Text preprocessing, sentiment score assignment	Time-series plots, sentiment over time	Assess sentiment in the context of identified key phrases using n-grams
3. Are there noticeable differences in the key terms used by various companies, and what might these differences indicate about their strategic priorities?	Comparative Text Analysis, N-gram Analysis	Standardization of text for comparative analysis	Heatmaps, comparative bar charts	Use n-grams to analyze strategic language patterns and differences between companies
4. How does sentiment within annual reports change over time, and can these changes be associated with specific business cycles or events?	Trend Analysis	Chronological ordering and segmentation of text data	Line graphs, trend lines	Examine changes in the frequency of n-grams related to sentiment to infer correlations with business cycles

Data Preparation Overview

- Data collection from EDGAR Database and Corporate Websites.
- Inclusion of the top companies, as of March 2024, in the Dow Jones Industrial Average (DJIA) by index weighting from 2019-2024 Fys
- 16 Companies; 5 most recent annual report filings. 80 documents

S/N	Company	Industry	Index Weighting
1	UnitedHealth Group	Healthcare	8.81%
2	Johnson & Johnson	Healthcare	2.70%
3	Merck & Co.	Healthcare	2.16%
4	Amgen	Healthcare	4.80%
5	Microsoft	Technology	6.83%
6	Apple	Technology	3.04%
7	Salesforce	Technology	5.04%
8	Intel Corporation	Technology	0.72%
9	Goldman Sachs	Financial Services	6.54%
10	Visa	Financial Services	4.76%
11	JPMorgan Chase	Financial Services	3.07%
12	American Express	Financial Services	3.64%
13	Home Depot	FMCG	6.23%
14	Walmart	FMCG	2.93%
15	Procter & Gamble	FMCG	2.69%
16	Nike	FMCG	1.75%

Text Processing & Data Transformation

- Imported the annual reports into a corpus: **DJ-ARC**
- Stored this corpus in a `tibble`.
- Cleaning text by removing punctuation, numbers, and converting to lowercase.
- Tokenization into words and sentences.
- Standardization for consistent formatting.
- Description of calculating Term Frequency (TF) and TF-IDF.
- N-grams development.

Custom Stopwords

```
custom_stopwords <- c(
  "financial", "statement", "report", "year", "company", "business",
  "management", "operations", "performance", "results", "objective",
  "strategy", "risk", "opportunity", "outlook", "significantly",
  "approximately", "primarily", "including", "regarding", "concerning",
  "due to", "pursuant", "accordance", "thereof", "therein", "hereby",
  "hereto", "hereunder", "usd", "ebitda", "gaap", "qoq", "yoy", "fy",
  "qtr", "one", "two", "first", "second", "third", "quarter", "annual",
  "monthly", "weekly", "day", "month", "year", "law", "regulation",
  "section", "act", "legal", "compliance", "regulatory", "filings",
  "securities", "exchange", "commission", "corporation", "incorporated",
  "plc", "llc", "ltd", "group", "holdings", "united states", "us", "usa",
  "america", "north america", "international", "global", "worldwide", "%",

   # Updated company names as stopwords with variations
  "unitedhealth group", "unitedhealth", "uhg",
  "johnson & johnson", "johnson and johnson", "johnson", "j&j",
  "merck", "merck & co", "mrk", "amgen", "amgn", "microsoft", "msft", "apple",
  "aapl", "salesforce", "crm", "intel", "intel corporation", "intc",
  "goldman sachs", "goldman", "gs", "visa", "v", "jpmorgan chase", "jpmorgan",
  "jpm", "american express", "amex", "axp", "home depot", "hd", "walmart",
  "wmt", "procter & gamble", "pg", "p&g", "procter", "gamble", "nike", "nke"
)
```

Initial Analysis

Top terms per company per year; head()

A tibble: 6 × 7

doc_id <chr>	industry <chr>
GS_2019	Financial Service
GS_2019	Financial Service
GS_2020	Financial Service
GS_2019	Financial Service
GS_2023	Financial Service
GS_2020	Financial Service

6 rows

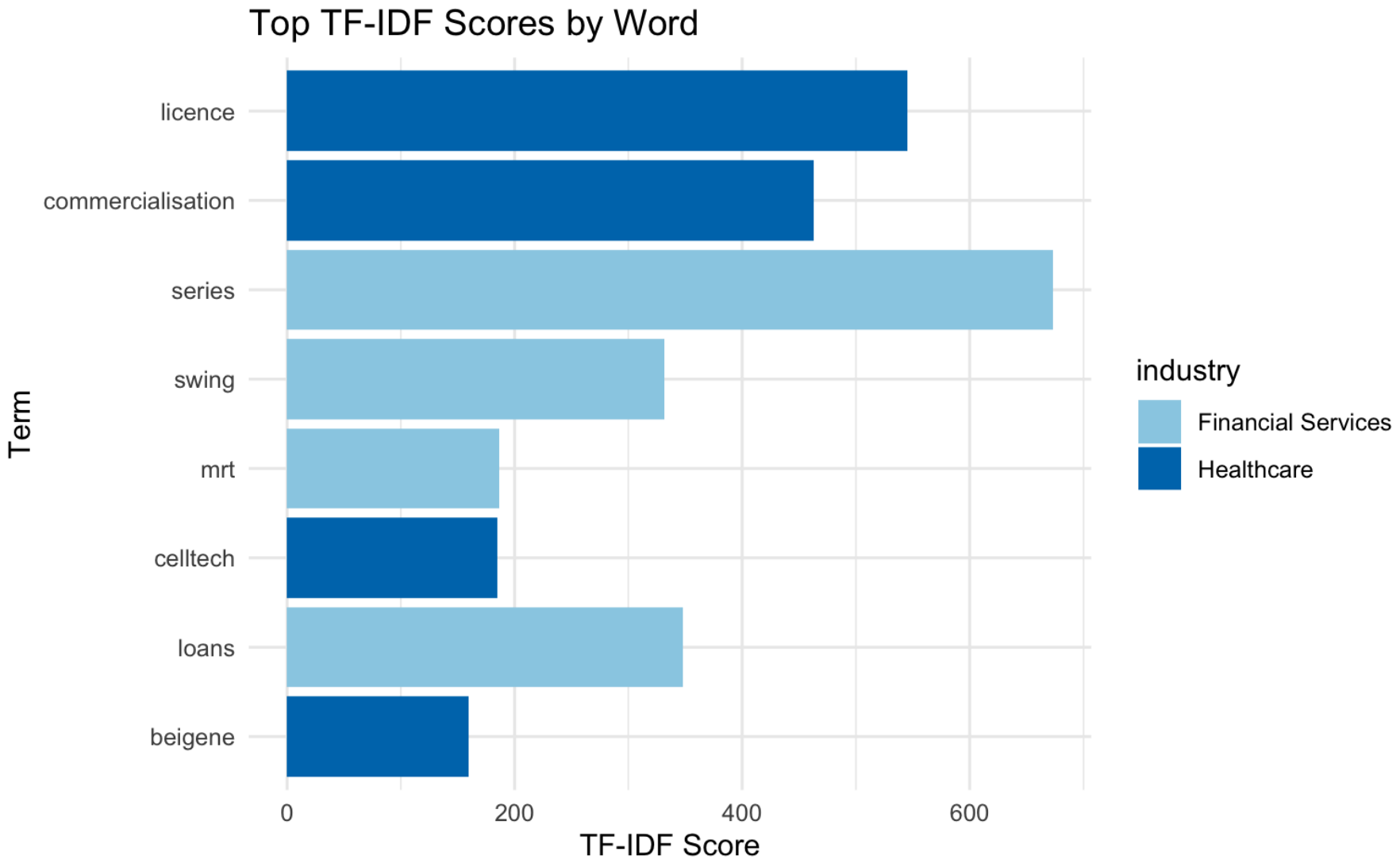
Top terms per company per year; tail()

A tibble: 6 × 7

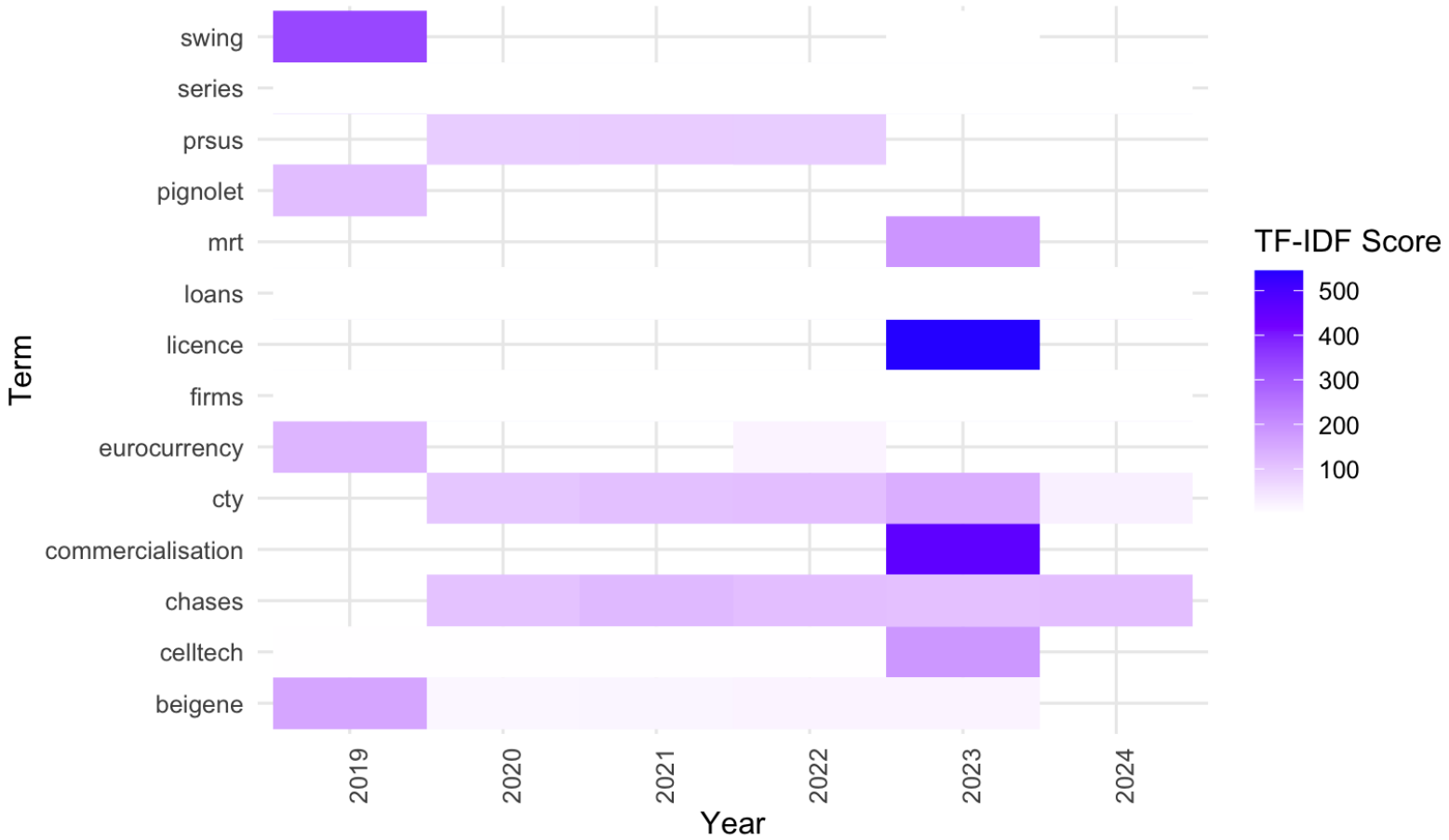
doc_id <chr>	industry <chr>	company_abbreviation <chr>	company_name <chr>	year <chr>	word <chr>	n <int>
WMT_2024	FMCG	WMT	Walmart	2024	yost	1
WMT_2024	FMCG	WMT	Walmart	2024	zambia	1
WMT_2024	FMCG	WMT	Walmart	2024	zappala	1
WMT_2024	FMCG	WMT	Walmart	2024	zdb	1
WMT_2024	FMCG	WMT	Walmart	2024	zip	1
WMT_2024	FMCG	WMT	Walmart	2024	znb	1

6 rows

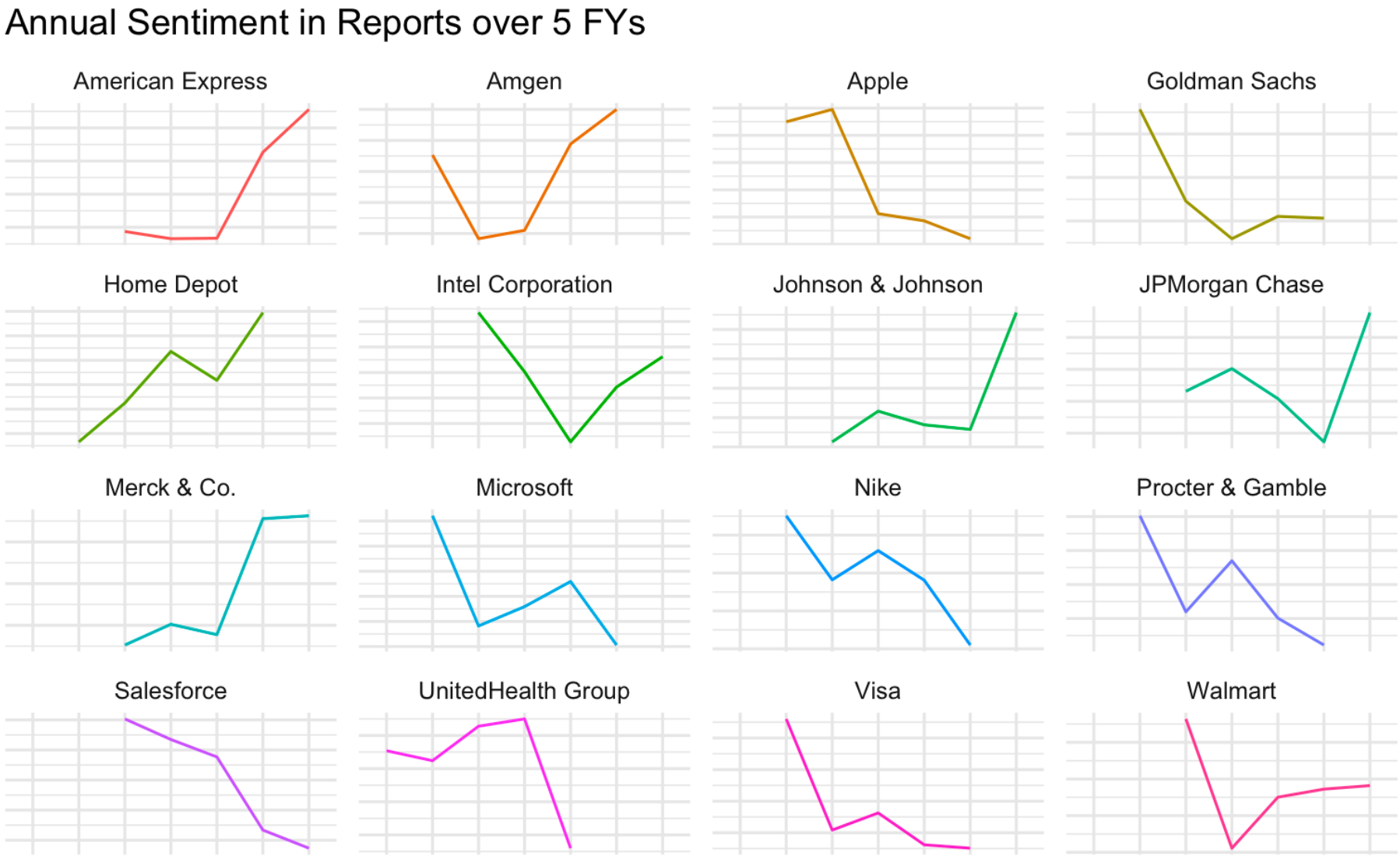
Research Question #1



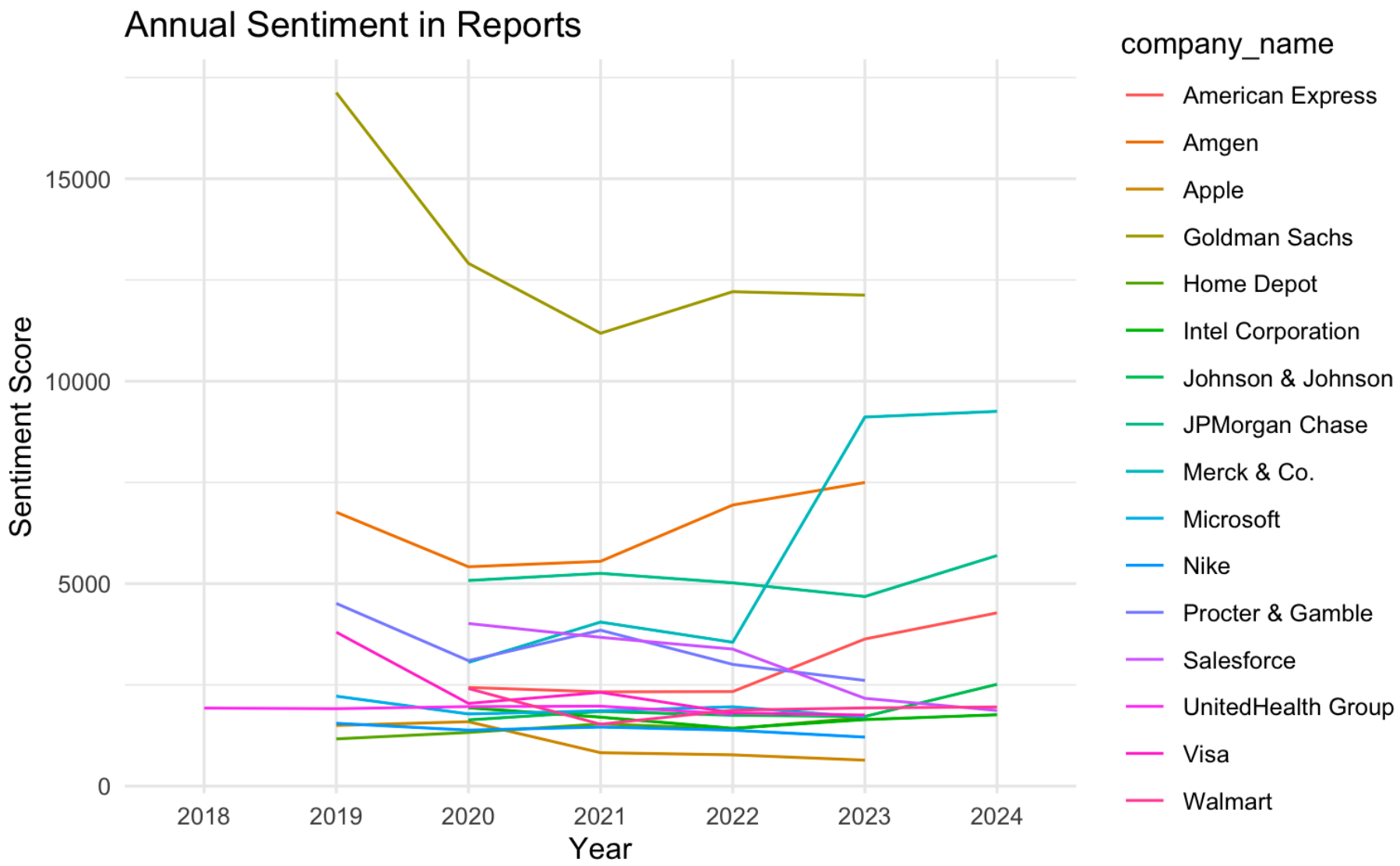
Top 10 TF-IDF Terms by Year



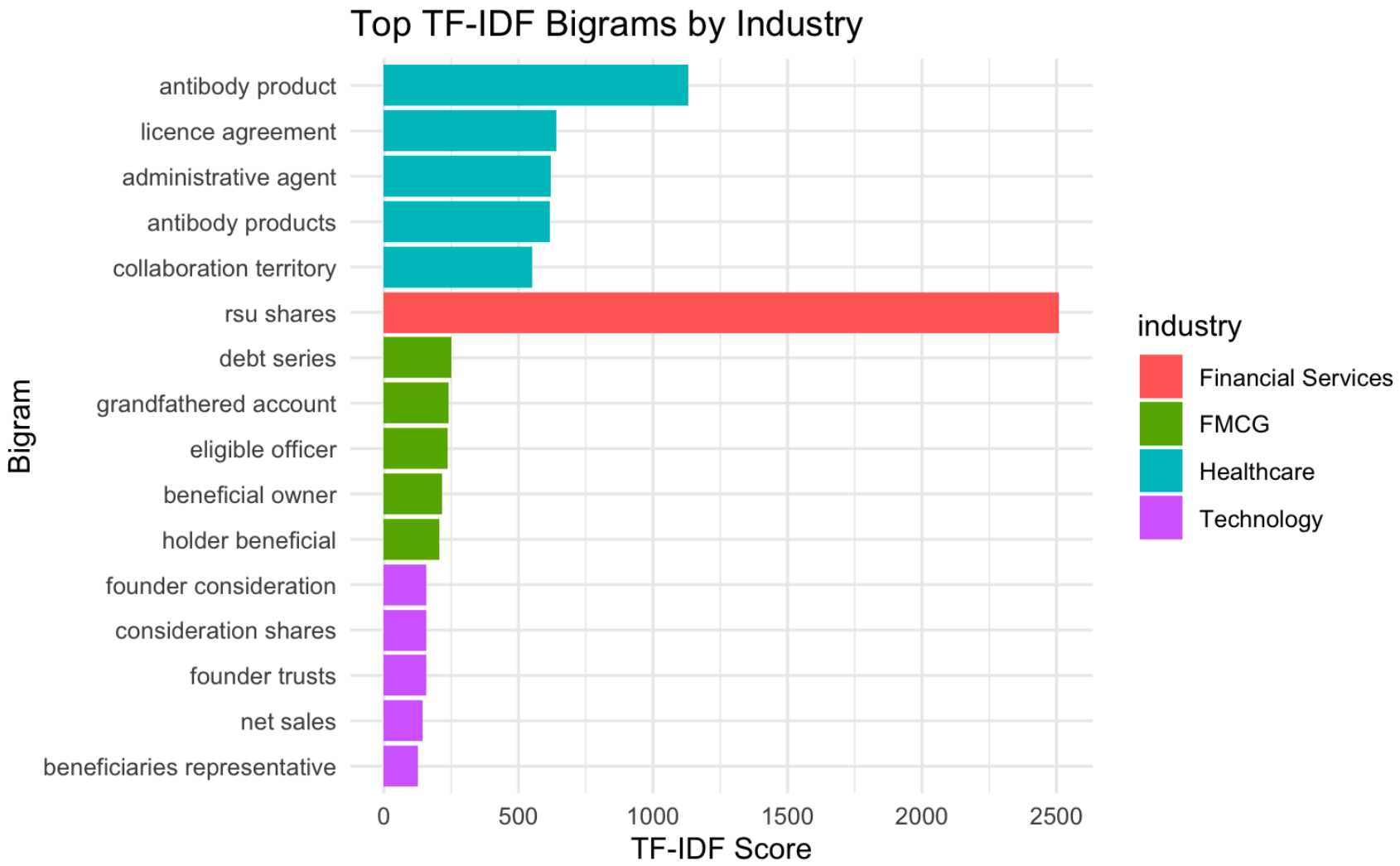
Research Question #2



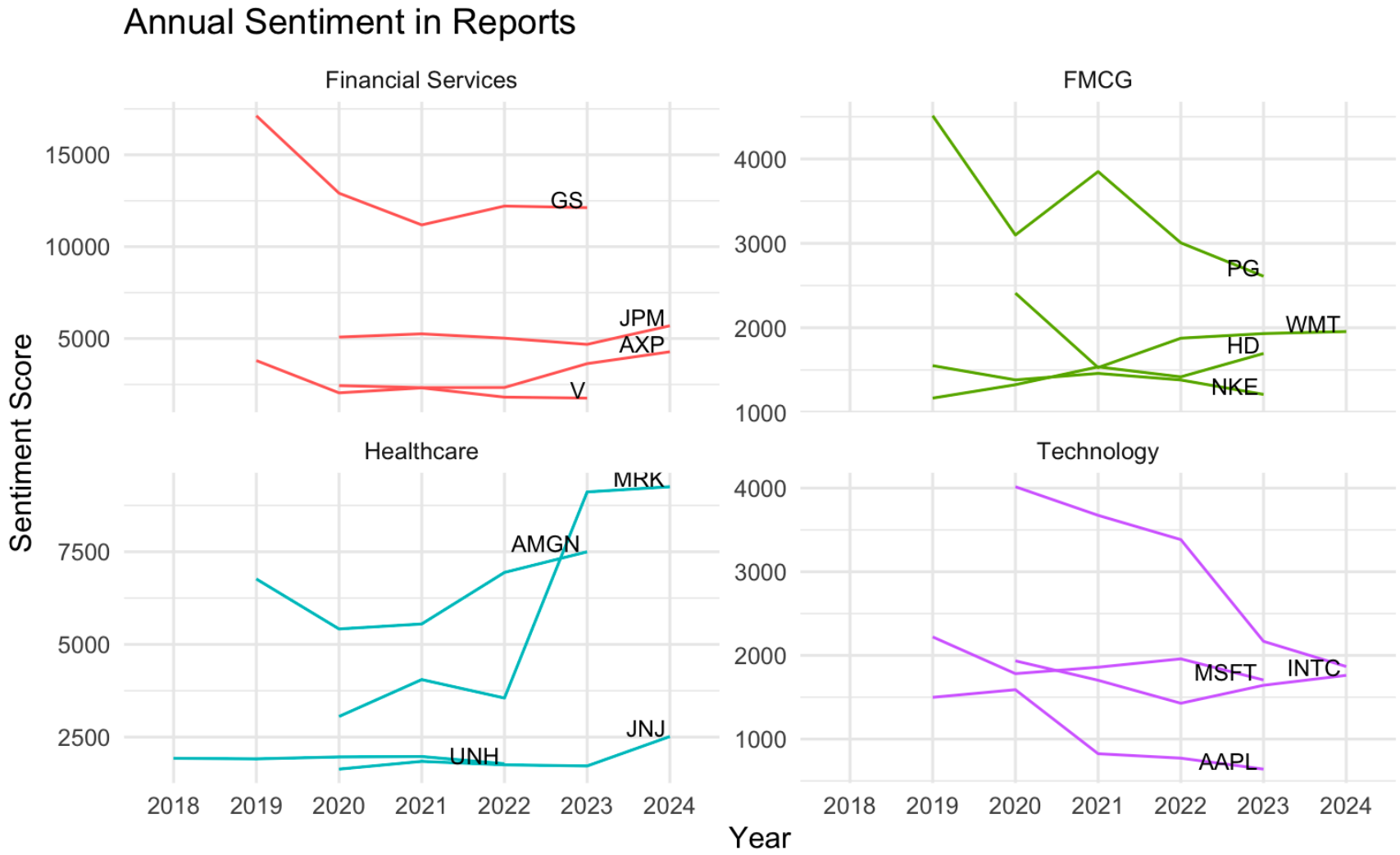
Research Question #2



Research Question #3



Research Question #4



Overall Insights

- Key term analysis indicated distinct areas of focus within industries, such as the emphasis on "commercialisation" in healthcare and "series" in financial services. Sentiment analysis illustrated how companies' tones fluctuated with industry conditions, with healthcare showing a generally positive sentiment trend and technology exhibiting greater volatility.
- The Term Frequency-Inverse Document Frequency (TF-IDF) analysis identified the terms "licence," "commercialisation," "series," and "swing," among others, as significant within the corpus. Specifically, "licence" and "commercialisation" were prominent in the healthcare industry, possibly indicating a strategic emphasis on product development and market access The term "series," mainly found in financial services, suggests a focus on investment products and financial instruments.
- The overall sentiment in the healthcare sector showed a notable upward trend, possibly reflecting positive developments or successful strategic initiatives. In contrast, the sentiment in the technology sector displayed a downward trend, which could be attributed to various factors, including market competition or technological disruptions.

Thank You