ESG Report Sentiment Analysis: Detecting Greenwashing and Industry Priorities

In [ ]:
```
#!pip install PyPDF2 nltk textblob pandas numpy matplotlib seaborn scikit-learn wordcloud
```

In [1]:
```
import os
import re
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from collections import Counter
import warnings
warnings.filterwarnings('ignore')

# For PDF processing
import PyPDF2

# For NLP
import nltk
from nltk.corpus import stopwords, wordnet
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.stem import WordNetLemmatizer

# For Sentiment Analysis
from textblob import TextBlob
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

# For TF-IDF
from sklearn.feature_extraction.text import TfidfVectorizer

# Downloading required NLTK data
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\sonali\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\sonali\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\sonali\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\sonali\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\sonali\AppData\Roaming\nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

Out[1]: True

In [2]:
```
def get_synonyms(word):
    """Grab synonyms from WordNet"""
    synonyms = set()
    for syn in wordnet.synsets(word):
        for lemma in syn.lemmas():
```

```
            synonym = lemma.name().replace('_', ' ').lower()
            synonyms.add(synonym)
    return synonyms

def expand_keywords_with_synonyms(keywords_list, max_synonyms_per_word=3):
    """Add synonyms to our keyword list"""
    expanded = set(keywords_list)
    for keyword in keywords_list:
        synonyms = get_synonyms(keyword)
        for syn in list(synonyms)[:max_synonyms_per_word]:
            if len(syn) > 2:  # skip really short words
                expanded.add(syn)
    return sorted(list(expanded))

# Starting keywords for each ESG pillar
# These are tailored to the 5 industries we're analyzing
BASE_ESG_KEYWORDS = {
    'Environmental': [
        # General environmental stuff
        'climate', 'carbon', 'emission', 'emissions', 'renewable', 'energy',
        'sustainability', 'environmental', 'waste', 'pollution', 'water',
        'greenhouse', 'ghg', 'biodiversity', 'ecosystem', 'conservation',
        'sustainable', 'ecological', 'footprint', 'recycling',

        # Oil & Gas
        'upstream', 'downstream', 'flaring', 'methane', 'scope 1', 'scope 2', 'scope 3',
        'carbon capture', 'ccs', 'ccus', 'net zero', 'transition', 'carbon intensity',
        'energy transition', 'low carbon', 'paris agreement', 'tcfd', 'spill',
        'contamination', 'remediation', 'produced water', 'fracking', 'offshore',

        # Fashion
        'textile', 'fabric', 'cotton', 'polyester', 'dyeing', 'tanning', 'leather',
        'microfiber', 'microplastic', 'chemical', 'toxic', 'wastewater', 'packaging',
        'circular fashion', 'resale', 'secondhand', 'repair', 'durability',
        'organic cotton', 'sustainable materials', 'recycled materials',

        # Agriculture
        'deforestation', 'land use', 'soil', 'pesticide', 'fertilizer', 'irrigation',
        'crop', 'farming', 'agriculture', 'agroforestry', 'regenerative',
        'palm oil', 'soy', 'cattle', 'livestock', 'monoculture', 'organic',
        'food waste', 'yield', 'drought', 'forest', 'habitat loss',

        # Pharma & Healthcare
        'pharmaceutical waste', 'drug disposal', 'medical waste', 'hazardous waste',
        'laboratory', 'clinical', 'packaging waste', 'green chemistry',
        'environmental footprint', 'facility', 'operations',
    ],

    'Social': [
        # General social terms
        'employee', 'employees', 'workforce', 'labor', 'worker', 'workers',
        'diversity', 'inclusion', 'equity', 'dei', 'equality', 'gender',
        'safety', 'health', 'wellbeing', 'training', 'development',
        'community', 'stakeholder', 'engagement', 'human rights',
        'discrimination', 'harassment', 'workplace', 'culture',

        # Fashion
        'supply chain', 'supplier', 'factory', 'garment worker', 'living wage',
        'fair labor', 'working conditions', 'forced labor', 'child labor',
        'freedom of association', 'collective bargaining', 'audit', 'inspection',
        'transparency', 'traceability', 'ethical sourcing', 'fair trade',

        # Oil & Gas
        'indigenous', 'indigenous rights', 'land rights', 'local communities',
        'displacement', 'consultation', 'free prior informed consent', 'fpic',
        'occupational safety', 'contractor', 'operational safety',
```

```
        # Agriculture
        'smallholder', 'farmer', 'farm worker', 'seasonal worker', 'migrant worker',
        'rural community', 'food security', 'nutrition', 'land rights',
        'fair price', 'cooperative', 'certification', 'standard',

        # Pharma & Healthcare
        'patient', 'patient access', 'affordability', 'pricing', 'drug pricing',
        'clinical trial', 'trial participant', 'informed consent', 'ethics committee',
        'access to medicine', 'essential medicine', 'global health', 'pandemic',
        'vaccine', 'rare disease', 'neglected disease', 'healthcare access',
        'physician', 'pharmacist', 'research ethics', 'bioethics',
        'patient care', 'patient safety', 'quality of care', 'medical staff',
        'nurse', 'burnout', 'staffing', 'emergency',
    ],

    'Governance': [
        # General governance
        'governance', 'board', 'director', 'directors', 'executive', 'ceo',
        'ethics', 'compliance', 'transparency', 'accountability', 'audit',
        'risk', 'management', 'shareholder', 'shareholders', 'integrity',
        'policy', 'policies', 'regulation', 'regulatory', 'oversight',
        'committee', 'disclosure', 'reporting', 'independent',

        # Ethics & corruption
        'corruption', 'bribery', 'anti-corruption', 'anti-bribery', 'whistleblower',
        'conflict of interest', 'code of conduct', 'ethical', 'misconduct',

        # ESG governance
        'esg committee', 'sustainability committee', 'materiality', 'framework',
        'gri', 'sasb', 'cdp', 'standards', 'assurance', 'verification',

        # Pharma specific
        'fda', 'ema', 'regulatory approval', 'pharmacovigilance', 'adverse event',
        'clinical governance', 'drug safety', 'quality control', 'gmp',
        'intellectual property', 'patent', 'r&d', 'research integrity',
        'marketing practices', 'physician payment', 'kickback',

        # Healthcare specific
        'hipaa', 'patient privacy', 'data protection', 'medical records',
        'malpractice', 'credentialing', 'accreditation', 'quality standard',

        # Financial governance
        'compensation', 'executive pay', 'bonus', 'incentive', 'stock option',
        'say on pay', 'proxy', 'voting', 'fiduciary',

        # Cybersecurity
        'cybersecurity', 'data privacy', 'data breach', 'gdpr', 'information security',
    ]
}

print("Expanding keywords with synonyms...\n")

# Add synonyms to each pillar
ESG_KEYWORDS = {}
for pillar, keywords in BASE_ESG_KEYWORDS.items():
    expanded = expand_keywords_with_synonyms(keywords, max_synonyms_per_word=2)
    ESG_KEYWORDS[pillar] = expanded
    print(f"{pillar}: {len(keywords)} → {len(expanded)} keywords")

# Multi-word phrases (WordNet can't handle these)
ADDITIONAL_PHRASES = {
    'Environmental': [
        'renewable energy', 'clean energy', 'net zero', 'decarbonization',
        'circular economy', 'resource efficiency', 'climate change',
        'carbon footprint', 'carbon neutral', 'carbon dioxide', 'co2',
```

```
        'ghg emissions', 'global warming', 'environmental impact',
        'life cycle assessment', 'cradle to cradle', 'zero waste',
        'water stewardship', 'water scarcity', 'air quality',
        'fast fashion', 'slow fashion', 'textile waste', 'fashion industry',
        'fossil fuel', 'natural gas', 'crude oil', 'oil spill', 'gas leak',
        'sustainable agriculture', 'climate smart agriculture', 'soil health',
        'crop rotation', 'precision agriculture', 'water management',
    ],
    'Social': [
        'human rights', 'customer satisfaction', 'supply chain', 'fair trade',
        'living wage', 'local communities', 'work life balance', 'employee engagement',
        'diversity and inclusion', 'pay equity', 'occupational health', 'labor rights',
        'community engagement', 'social responsibility', 'fair labor',
        'health and safety', 'mental health', 'employee wellbeing',
        'garment workers', 'factory conditions', 'supply chain transparency',
        'patient access', 'drug pricing', 'clinical trials', 'vaccine access',
        'health equity', 'healthcare workers', 'food security', 'smallholder farmers',
    ],
    'Governance': [
        'corporate governance', 'board of directors', 'risk management',
        'executive compensation', 'conflict of interest', 'code of conduct',
        'internal controls', 'compliance program', 'board independence',
        'shareholder rights', 'corporate ethics', 'whistleblower protection',
        'anti corruption', 'data privacy', 'cyber security',
        'esg reporting', 'sustainability reporting', 'materiality assessment',
        'drug safety', 'regulatory compliance', 'clinical trial ethics',
        'marketing practices', 'research integrity', 'patient privacy', 'medical ethics',
    ]
}

for pillar, phrases in ADDITIONAL_PHRASES.items():
    ESG_KEYWORDS[pillar].extend(phrases)
    ESG_KEYWORDS[pillar] = sorted(list(set(ESG_KEYWORDS[pillar])))

print(f"\nFinal counts:")
for pillar, keywords in ESG_KEYWORDS.items():
    print(f"  {pillar}: {len(keywords)} keywords")

print(f"\nSample keywords:")
for pillar, keywords in ESG_KEYWORDS.items():
    print(f"{pillar}: {', '.join(keywords[:15])}...")

# Vague language that might signal greenwashing
GREENWASHING_INDICATORS = [
    'committed', 'commitment', 'dedication', 'dedicated', 'passionate', 'leading',
    'strive', 'striving', 'world-class', 'best-in-class', 'innovative', 'excellence',
    'endeavor', 'endeavoring', 'working towards', 'aiming', 'planning', 'intend',
    'aspire', 'aspiring', 'believe', 'proud', 'excited', 'promising', 'exploring',
    'journey', 'vision', 'ambition', 'passionate about', 'hope', 'hoping', 'desire',
    'seek', 'seeking', 'continue to', 'ongoing', 'long term', 'future',
    'aware', 'recognize', 'understand', 'acknowledge', 'considering',
]

GREENWASHING_INDICATORS = expand_keywords_with_synonyms(GREENWASHING_INDICATORS, max_synor
print(f"\nGreenwashing indicators: {len(GREENWASHING_INDICATORS)} terms")

# Concrete action words (the opposite of greenwashing)
SUBSTANTIVE_WORDS = [
    'achieved', 'reduced', 'increased', 'implemented', 'completed', 'delivered',
    'measured', 'reported', 'certified', 'audited', 'verified', 'reached',
    'target', 'goal', 'metric', 'data', 'performance', 'result', 'outcome',
    'baseline', 'benchmark', 'kpi', 'indicator', 'quantified', 'tracked',
    'invested', 'spent', 'allocated', 'million', 'billion', 'percent', 'percentage',
    'launched', 'established', 'created', 'installed', 'deployed', 'executed',
    'eliminated', 'decreased', 'improved', 'upgraded', 'retrofitted',
    'validated', 'assessed', 'monitored', 'disclosed', 'published',
```

```
]

SUBSTANTIVE_WORDS = expand_keywords_with_synonyms(SUBSTANTIVE_WORDS, max_synonyms_per_word
print(f"Substantive action words: {len(SUBSTANTIVE_WORDS)} terms")

# Save everything to file for reference
with open('expanded_esg_keywords.txt', 'w') as f:
    for pillar, keywords in ESG_KEYWORDS.items():
        f.write(f"\n{pillar.upper()} ({len(keywords)} keywords)\n")
        f.write("="*60 + "\n")
        f.write(', '.join(keywords) + '\n')

    f.write(f"\nGREENWASHING INDICATORS ({len(GREENWASHING_INDICATORS)} terms)\n")
    f.write("="*60 + "\n")
    f.write(', '.join(GREENWASHING_INDICATORS) + '\n')

    f.write(f"\nSUBSTANTIVE WORDS ({len(SUBSTANTIVE_WORDS)} terms)\n")
    f.write("="*60 + "\n")
    f.write(', '.join(SUBSTANTIVE_WORDS) + '\n')

print(f"\nKeywords saved to 'expanded_esg_keywords.txt'")
```

```
Expanding keywords with synonyms...

Environmental: 97 → 161 keywords
Social: 95 → 153 keywords
Governance: 82 → 142 keywords

Final counts:
  Environmental: 195 keywords
  Social: 173 keywords
  Governance: 160 keywords

Sample keywords:
Environmental: aflare, agribusiness, agricultural, agriculture, agriculture department, ag
roforestry, air quality, biodiversity, bionomic, bos taurus, bronze, browse, carbon, carbo
n capture, carbon copy...
Social: access to medicine, acculturation, affordability, apothecary, audit, autochthonou
s, base hit, battle, bioethics, biotic community, booking, breeding, burnout, certificatio
n, child labor...
Governance: abidance, accountability, accreditation, administration, administrator, advers
e event, answerableness, anti corruption, anti-bribery, anti-corruption, assurance, audit,
balloting, board, board independence...

Greenwashing indicators: 97 terms
Substantive action words: 121 terms

Keywords saved to 'expanded_esg_keywords.txt'
```

In [3]:

```
# Extract text from PDF file
def extract_text_from_pdf(pdf_path):

    text = ""
    try:
        with open(pdf_path, 'rb') as file:
            pdf_reader = PyPDF2.PdfReader(file)
            num_pages = len(pdf_reader.pages)

            for page_num in range(num_pages):
                page = pdf_reader.pages[page_num]
                page_text = page.extract_text()
                if page_text:
                    text += page_text + " "

        return text, num_pages
```

```
    except Exception as e:
        print(f"Error reading {pdf_path}: {str(e)}")
        return "", 0

# Clean up extracted text
def clean_text(text):
    # Fix spacing issues
    text = re.sub(r'\s+', ' ', text)
    # Keep letters, numbers, periods, commas, percent signs
    text = re.sub(r'[^\w\s\.\,\%]', ' ', text)
    # Remove standalone numbers (but keep percentages)
    text = re.sub(r'\b\d+\b', '', text)
    return text.strip()
```

In [4]:
```
# Analyze sentiment using VADER and TextBlob
def analyze_sentiment(text):
    # VADER sentiment scores
    vader = SentimentIntensityAnalyzer()
    vader_scores = vader.polarity_scores(text)

    # TextBlob sentiment
    blob = TextBlob(text)
    textblob_polarity = blob.sentiment.polarity
    textblob_subjectivity = blob.sentiment.subjectivity

    # Sentence-level analysis
    sentences = sent_tokenize(text)

    if len(sentences) > 0:
        sentence_sentiments = [TextBlob(sent).sentiment.polarity for sent in sentences]
        positive_sentences = sum(1 for s in sentence_sentiments if s > 0.1)
        negative_sentences = sum(1 for s in sentence_sentiments if s < -0.1)
        neutral_sentences = len(sentences) - positive_sentences - negative_sentences
        avg_sentiment = np.mean(sentence_sentiments)
        sentiment_std = np.std(sentence_sentiments)
    else:
        sentence_sentiments = [0]
        positive_sentences = 0
        negative_sentences = 0
        neutral_sentences = 0
        avg_sentiment = 0
        sentiment_std = 0

    return {
        'vader_compound': vader_scores['compound'],
        'vader_positive': vader_scores['pos'],
        'vader_negative': vader_scores['neg'],
        'vader_neutral': vader_scores['neu'],
        'textblob_polarity': textblob_polarity,
        'textblob_subjectivity': textblob_subjectivity,
        'avg_sentence_sentiment': avg_sentiment,
        'sentence_sentiment_std': sentiment_std,
        'total_sentences': len(sentences),
        'positive_sentences': positive_sentences,
        'negative_sentences': negative_sentences,
        'neutral_sentences': neutral_sentences,
        'positive_sentences_pct': (positive_sentences / len(sentences) * 100) if len(sente
    }
```

In [5]:
```
def detect_greenwashing(text):
    """Detect greenwashing by comparing aspirational vs concrete language"""

    text_lower = text.lower()
    words = word_tokenize(text_lower)
    total_words = len(words)
```

```
    if total_words == 0:
        return {
            'greenwashing_indicators': 0,
            'substantive_words': 0,
            'greenwashing_density': 0,
            'substantive_density': 0,
            'greenwashing_ratio': 0,
            'risk_level': 'Unknown',
            'risk_score': 0
        }

    # Count vague/aspirational language
    greenwashing_count = 0
    for indicator in GREENWASHING_INDICATORS:
        greenwashing_count += text_lower.count(indicator)

    # Count concrete actions/metrics
    substantive_count = 0
    for word in SUBSTANTIVE_WORDS:
        substantive_count += text_lower.count(word)

    # Calculate density per 1000 words
    greenwashing_density = (greenwashing_count / total_words) * 1000
    substantive_density = (substantive_count / total_words) * 1000

    # Calculate ratio: high ratio = more fluff than substance
    if substantive_count > 0:
        greenwashing_ratio = greenwashing_count / substantive_count
    else:
        greenwashing_ratio = greenwashing_count if greenwashing_count > 0 else 0

    # Risk assessment
    risk_score = (greenwashing_ratio * 0.6) + (greenwashing_density * 0.4)

    if greenwashing_ratio > 1.5 or greenwashing_density > 15:
        risk_level = 'High'
    elif greenwashing_ratio > 0.8 or greenwashing_density > 8:
        risk_level = 'Medium'
    else:
        risk_level = 'Low'

    return {
        'greenwashing_indicators': greenwashing_count,
        'substantive_words': substantive_count,
        'greenwashing_density': round(greenwashing_density, 2),
        'substantive_density': round(substantive_density, 2),
        'greenwashing_ratio': round(greenwashing_ratio, 2),
        'risk_level': risk_level,
        'risk_score': round(risk_score, 2)
    }
```

In [6]:

```
# Calculate which ESG pillars the company focuses on
def calculate_esg_importance(text, company_name="Company"):

    text_lower = text.lower()

    # Count keyword mentions for each pillar
    esg_counts = {}
    for pillar, keywords in ESG_KEYWORDS.items():
        count = 0
        for keyword in keywords:
            count += text_lower.count(keyword.lower())
        esg_counts[pillar] = count

    total_keywords = sum(esg_counts.values())
```

```
    if total_keywords == 0:
        return {
            'Environmental_count': 0,
            'Social_count': 0,
            'Governance_count': 0,
            'Environmental_pct': 0,
            'Social_pct': 0,
            'Governance_pct': 0,
            'dominant_pillar': 'None',
            'total_esg_keywords': 0
        }

    # Calculate percentages
    env_pct = (esg_counts['Environmental'] / total_keywords) * 100
    soc_pct = (esg_counts['Social'] / total_keywords) * 100
    gov_pct = (esg_counts['Governance'] / total_keywords) * 100

    # Find dominant pillar
    dominant_pillar = max(esg_counts.items(), key=lambda x: x[1])[0]

    return {
        'Environmental_count': esg_counts['Environmental'],
        'Social_count': esg_counts['Social'],
        'Governance_count': esg_counts['Governance'],
        'Environmental_pct': round(env_pct, 2),
        'Social_pct': round(soc_pct, 2),
        'Governance_pct': round(gov_pct, 2),
        'dominant_pillar': dominant_pillar,
        'total_esg_keywords': total_keywords
    }
```

In [7]:

```
# Using TF-IDF to compare ESG focus across companies
def tfidf_esg_analysis(texts_dict):

    companies = list(texts_dict.keys())
    documents = list(texts_dict.values())

    results = []

    for pillar, keywords in ESG_KEYWORDS.items():
        # TF-IDF with ESG keywords as vocabulary
        vectorizer = TfidfVectorizer(
            vocabulary=keywords,
            lowercase=True,
            token_pattern=r'\b\w+\b'
        )

        try:
            tfidf_matrix = vectorizer.fit_transform(documents)

            # Get scores for each company
            for idx, company in enumerate(companies):
                doc_scores = tfidf_matrix[idx].toarray().flatten()
                mean_tfidf = np.mean(doc_scores) if len(doc_scores) > 0 else 0
                max_tfidf = np.max(doc_scores) if len(doc_scores) > 0 else 0

                results.append({
                    'company': company,
                    'pillar': pillar,
                    'tfidf_mean': round(mean_tfidf, 4),
                    'tfidf_max': round(max_tfidf, 4)
                })

        except Exception as e:
            print(f"TF-IDF issue for {pillar}: {e}")
```

```
            for company in companies:
                results.append({
                    'company': company,
                    'pillar': pillar,
                    'tfidf_mean': 0,
                    'tfidf_max': 0
                })

    return pd.DataFrame(results)
```

In [8]:

```
def analyze_single_esg_report(pdf_path, company_name, industry, controversy_level="Low"):
    """Run complete analysis on one ESG report"""

    print(f"\n{'='*60}")
    print(f"Analyzing: {company_name}")
    print(f"Industry: {industry} | Controversy: {controversy_level}")
    print(f"{'='*60}")

    # Extract text
    text, num_pages = extract_text_from_pdf(pdf_path)
    if not text or len(text) < 100:
        print("Failed to extract text")
        return None

    clean_text_content = clean_text(text)
    word_count = len(word_tokenize(clean_text_content))

    print(f"Extracted {len(text):,} characters from {num_pages} pages")
    print(f"Word count: {word_count:,}")

    # Sentiment analysis
    sentiment = analyze_sentiment(clean_text_content)
    print(f"\nSentiment:")
    print(f"  VADER: {sentiment['vader_compound']:.3f}")
    print(f"  TextBlob: {sentiment['textblob_polarity']:.3f}")
    print(f"  Positive sentences: {sentiment['positive_sentences_pct']:.1f}%")

    # Greenwashing detection
    greenwashing = detect_greenwashing(clean_text_content)
    print(f"\nGreenwashing:")
    print(f"  Risk: {greenwashing['risk_level']}")
    print(f"  Ratio: {greenwashing['greenwashing_ratio']:.2f}")
    print(f"  Aspirational words: {greenwashing['greenwashing_indicators']}")
    print(f"  Substantive words: {greenwashing['substantive_words']}")

    # ESG pillar analysis
    esg_importance = calculate_esg_importance(clean_text_content, company_name)
    print(f"\nESG Focus:")
    print(f"  Environmental: {esg_importance['Environmental_pct']:.1f}%")
    print(f"  Social: {esg_importance['Social_pct']:.1f}%")
    print(f"  Governance: {esg_importance['Governance_pct']:.1f}%")
    print(f"  Dominant: {esg_importance['dominant_pillar']}")

    # Compile results
    results = {
        'company_name': company_name,
        'industry': industry,
        'controversy_level': controversy_level,
        'num_pages': num_pages,
        'word_count': word_count,
        'text_content': clean_text_content,
    }

    results.update(sentiment)
    results.update(greenwashing)
    results.update(esg_importance)
```

```
    print(f"\nAnalysis complete for {company_name}\n")

    return results
```

In [9]:
```
# Analyze all ESG reports for one industry
def analyze_industry_reports(pdf_folder, industry_name, company_info):

    print(f"\n{'#'*60}")
    print(f"INDUSTRY: {industry_name}")
    print(f"{'#'*60}")

    all_results = []
    texts_for_tfidf = {}

    for filename, (company_name, controversy) in company_info.items():
        pdf_path = os.path.join(pdf_folder, filename)

        if not os.path.exists(pdf_path):
            print(f"File not found: {pdf_path}")
            continue

        # Analyze report
        result = analyze_single_esg_report(pdf_path, company_name, industry_name, controve

        if result:
            all_results.append(result)
            texts_for_tfidf[company_name] = result['text_content']

    # Create results dataframe
    df = pd.DataFrame(all_results)

    # TF-IDF analysis across companies
    if len(texts_for_tfidf) > 1:
        print(f"\nRunning TF-IDF analysis across {len(texts_for_tfidf)} companies...")
        tfidf_df = tfidf_esg_analysis(texts_for_tfidf)
    else:
        tfidf_df = pd.DataFrame()

    print(f"\n{'='*60}")
    print(f"Industry analysis complete: {industry_name}")
    print(f"Companies analyzed: {len(all_results)}")
    print(f"{'='*60}\n")

    return df, tfidf_df
```

In [10]:
```
def analyze_and_display_industry(industry_name, pdf_folder, company_info, save_excel=True

    # Show configuration
    print(f"\n{'='*60}")
    print(f"ANALYZING: {industry_name}")
    print(f"{'='*60}")
    print(f"PDF folder: {pdf_folder}")
    print(f"Companies: {len(company_info)}")
    print(f"\nCompanies to analyze:")
    for filename, (company, controversy) in company_info.items():
        print(f"  - {company} ({controversy} controversy)")

    # Run analysis
    results_df, tfidf_df = analyze_industry_reports(pdf_folder, industry_name, company_int

    # Display results
    excel_file = None
    if not results_df.empty:
        print("\n" + "="*60)
```

```
        print("RESULTS PREVIEW")
        print("="*60)
        display(results_df.head())

        print("\nKey Metrics:")
        key_cols = ['company_name', 'controversy_level', 'vader_compound',
                    'risk_level', 'greenwashing_ratio', 'dominant_pillar']
        available_cols = [col for col in key_cols if col in results_df.columns]
        display(results_df[available_cols])

        # Save to Excel
        if save_excel:
            from datetime import datetime
            timestamp = datetime.now().strftime("%Y%m%d_%H%M")
            excel_file = f"{industry_name.replace(' ', '_')}_ESG_Analysis_{timestamp}.xlsx

            print(f"\nSaving to Excel: {excel_file}")

            with pd.ExcelWriter(excel_file, engine='openpyxl') as writer:
                # Full results
                results_df.to_excel(writer, sheet_name='Full_Results', index=False)

                # Summary
                summary_cols = [
                    'company_name', 'controversy_level', 'num_pages', 'word_count',
                    'vader_compound', 'textblob_polarity', 'textblob_subjectivity',
                    'greenwashing_ratio', 'risk_level', 'risk_score',
                    'Environmental_pct', 'Social_pct', 'Governance_pct', 'dominant_pillar
                ]
                available_summary = [col for col in summary_cols if col in results_df.colu
                summary_df = results_df[available_summary].copy()
                summary_df.to_excel(writer, sheet_name='Summary', index=False)

                # TF-IDF
                if not tfidf_df.empty:
                    tfidf_df.to_excel(writer, sheet_name='TFIDF_Analysis', index=False)
                    tfidf_pivot = tfidf_df.pivot(index='company', columns='pillar', values
                    tfidf_pivot.to_excel(writer, sheet_name='TFIDF_Pivot')

                # Greenwashing ranking
                greenwash_cols = ['company_name', 'controversy_level', 'greenwashing_ratio
                                  'risk_level', 'risk_score']
                available_gw = [col for col in greenwash_cols if col in results_df.columns
                greenwash_rank = results_df[available_gw].copy()
                greenwash_rank = greenwash_rank.sort_values('greenwashing_ratio', ascendin
                greenwash_rank.to_excel(writer, sheet_name='Greenwashing_Ranking', index=F

                # Sentiment ranking
                sentiment_cols = ['company_name', 'controversy_level', 'vader_compound',
                available_sent = [col for col in sentiment_cols if col in results_df.colum
                sentiment_rank = results_df[available_sent].copy()
                sentiment_rank = sentiment_rank.sort_values('vader_compound', ascending=Fa
                sentiment_rank.to_excel(writer, sheet_name='Sentiment_Ranking', index=Fals

                # ESG pillars
                esg_cols = ['company_name', 'Environmental_pct', 'Social_pct',
                            'Governance_pct', 'dominant_pillar']
                available_esg = [col for col in esg_cols if col in results_df.columns]
                esg_comparison = results_df[available_esg].copy()
                esg_comparison.to_excel(writer, sheet_name='ESG_Pillars', index=False)

                # Controversy analysis
                if 'controversy_level' in results_df.columns:
                    controversy_analysis = results_df.groupby('controversy_level').agg({
                        'greenwashing_ratio': ['mean', 'min', 'max'],
                        'vader_compound': ['mean', 'min', 'max']
```

```
                    }).round(3)
                    controversy_analysis.to_excel(writer, sheet_name='Controversy_Analysis

            print(f"✓ Saved to: {excel_file}")
    else:
        print("\n⚠️ No results generated")

    return results_df, tfidf_df, excel_file


# Base path
BASE_PATH = r"C:\Users\sonali\OneDrive\Desktop\Trimester I\AN6002 Analytics and ML in Busi

# All industries configuration
INDUSTRIES_CONFIG = {
    "Fashion Retail": {
        "folder": f"{BASE_PATH}\\Fashion Retail",
        "companies": {
            "Aritzia.pdf": ("Aritzia", "Low"),
            "FastRetailing.pdf": ("Fast Retailing", "Medium"),
            "GAPInc.pdf": ("GAP Inc", "Medium"),
            "H&M.pdf": ("H&M", "High"),
            "Inditex.pdf": ("Inditex", "Medium"),
            "Levis.pdf": ("Levi's", "Low"),
            "LMVH.pdf": ("LVMH", "Low"),
            "Lululemon.pdf": ("Lululemon", "Low"),
            "M&S.pdf": ("Marks & Spencer", "Low"),
            "Patagonia.pdf": ("Patagonia", "Low"),
            "Prada.pdf": ("Prada", "Low"),
            "RalphLauren.pdf": ("Ralph Lauren", "Low"),
            "Shein.pdf": ("Shein", "High"),
            "TJX.pdf": ("TJX Companies", "Medium"),
            "VSCo.pdf": ("Victoria's Secret", "Medium"),
        }
    },

    "Healthcare": {
        "folder": f"{BASE_PATH}\\Healthcare",
        "companies": {
            "Cardinal.pdf": ("Cardinal Health", "Medium"),
            "Cencora.pdf": ("Cencora", "Low"),
            "Centene.pdf": ("Centene", "Medium"),
            "CignaGroup.pdf": ("The Cigna Group", "Low"),
            "CVSHealth.pdf": ("CVS Health", "Medium"),
            "ELV.pdf": ("Elevance Health", "Low"),
            "Humana.pdf": ("Humana", "Low"),
            "McKesson.pdf": ("McKesson", "Low"),
            "UHG.pdf": ("UnitedHealth Group", "Medium"),
            "WBA.pdf": ("Walgreens Boots Alliance", "Medium"),
        }
    },

    "Oil & Gas": {
        "folder": f"{BASE_PATH}\\Oil & Gas",
        "companies": {
            "Chevron.pdf": ("Chevron", "High"),
            "Enbridge.pdf": ("Enbridge", "Medium"),
            "ExxonMobil.pdf": ("ExxonMobil", "High"),
            "IndianOil.pdf": ("Indian Oil", "High"),
            "Marathon.pdf": ("Marathon Petroleum", "High"),
            "Occidental.pdf": ("Occidental Petroleum", "High"),
            "SaudiAramco.pdf": ("Saudi Aramco", "High"),
            "Shell.pdf": ("Shell", "High"),
            "SouthernCompany.pdf": ("Southern Company", "Medium"),
        }
    },
```

```
    "Agriculture": {
        "folder": f"{BASE_PATH}\\Agriculture",
        "companies": {
            "Astra Agro Lestari (ID).pdf": ("Astra Agro Lestari", "High"),
            "Cargill (US).pdf": ("Cargill", "High"),
            "COFCO (CN).pdf": ("COFCO", "Medium"),
            "DCM Shriram (IN).pdf": ("DCM Shriram", "Medium"),
            "Golden Agri-Resources (SG).pdf": ("Golden Agri-Resources", "High"),
            "GrainCorp (AU).pdf": ("GrainCorp", "Low"),
            "Kuala Lumpur Kepong (MY).pdf": ("Kuala Lumpur Kepong", "High"),
            "KWS (DE).pdf": ("KWS", "Low"),
            "M.P. Evans (UK).pdf": ("M.P. Evans", "High"),
            "Sakata (JP).pdf": ("Sakata", "Low"),
            "Tessenderlo Group (BE).pdf": ("Tessenderlo Group", "Medium"),
        }
    },

    "Pharma": {
        "folder": f"{BASE_PATH}\\Pharma",
        "companies": {
            "abbvie.pdf": ("AbbVie", "Medium"),
            "Agios.pdf": ("Agios Pharmaceuticals", "Low"),
            "AstraZeneca.pdf": ("AstraZeneca", "Low"),
            "Bristol Myers Squibb.pdf": ("Bristol Myers Squibb", "Low"),
            "Daiichi Sankyo.pdf": ("Daiichi Sankyo", "Low"),
            "GSK.pdf": ("GSK", "Medium"),
            "Merck.pdf": ("Merck", "Low"),
            "Pfizer_2023.pdf": ("Pfizer", "Medium"),
            "Roche.pdf": ("Roche", "Low"),
            "Takeda.pdf": ("Takeda", "Low"),
        }
    }
}
```

In [11]:

```
# Analyze all industries and save each to Excel
all_results = {}

for industry_name, config in INDUSTRIES_CONFIG.items():
    results_df, tfidf_df, excel_file = analyze_and_display_industry(
        industry_name,
        config["folder"],
        config["companies"],
        save_excel=True
    )

    if results_df is not None and not results_df.empty:
        all_results[industry_name] = {
            'results': results_df,
            'tfidf': tfidf_df,
            'excel_file': excel_file
        }

# Summary of all analyses
print(f"\n{'='*70}")
print("ALL INDUSTRIES COMPLETE")
print(f"{'='*70}")
for industry, data in all_results.items():
    print(f"✓ {industry}: {len(data['results'])} companies analyzed")
    print(f"  Excel file: {data['excel_file']}")

# Create combined master file
if all_results:
    from datetime import datetime
    timestamp = datetime.now().strftime("%Y%m%d_%H%M")
    master_file = f"ALL_INDUSTRIES_ESG_Analysis_{timestamp}.xlsx"
```

```
combined_df = pd.concat([data['results'] for data in all_results.values()], ignore_inc

print(f"\nCreating master file: {master_file}")

with pd.ExcelWriter(master_file, engine='openpyxl') as writer:
    # All industries combined
    combined_df.to_excel(writer, sheet_name='All_Industries', index=False)

    # Summary by industry
    summary_cols = ['industry', 'company_name', 'controversy_level', 'vader_compound',
                    'greenwashing_ratio', 'risk_level', 'Environmental_pct',
                    'Social_pct', 'Governance_pct', 'dominant_pillar']
    available = [col for col in summary_cols if col in combined_df.columns]
    combined_df[available].to_excel(writer, sheet_name='Summary', index=False)

    # Industry averages
    industry_stats = combined_df.groupby('industry').agg({
        'vader_compound': 'mean',
        'greenwashing_ratio': 'mean',
        'Environmental_pct': 'mean',
        'Social_pct': 'mean',
        'Governance_pct': 'mean'
    }).round(2)
    industry_stats.to_excel(writer, sheet_name='Industry_Averages')

    # Cross-industry greenwashing ranking
    greenwash_cols = ['industry', 'company_name', 'controversy_level',
                      'greenwashing_ratio', 'risk_level']
    available_gw = [col for col in greenwash_cols if col in combined_df.columns]
    all_greenwash = combined_df[available_gw].sort_values('greenwashing_ratio', ascend
    all_greenwash.to_excel(writer, sheet_name='Cross_Industry_Ranking', index=False)

    # Controversy vs Greenwashing
    if 'controversy_level' in combined_df.columns:
        controversy_green = combined_df.groupby(['industry', 'controversy_level']).agg
            'greenwashing_ratio': 'mean',
            'vader_compound': 'mean'
        }).round(3)
        controversy_green.to_excel(writer, sheet_name='Controversy_vs_Greenwash')

    # ESG pillar dominance by industry
    pillar_by_industry = pd.crosstab(combined_df['industry'], combined_df['dominant_pi
    pillar_by_industry.to_excel(writer, sheet_name='Pillar_by_Industry')

print(f"✓ Master file saved: {master_file}")
print(f"\n{'='*70}")
print("ANALYSIS COMPLETE!")
print(f"{'='*70}")
print(f"Total companies analyzed: {len(combined_df)}")
print(f"Individual industry files: {len(all_results)}")
print(f"Combined master file: {master_file}")
```

```
============================================================
ANALYZING: Fashion Retail
============================================================
PDF folder: C:\Users\sonali\OneDrive\Desktop\Trimester I\AN6002 Analytics and ML in Busine
ss\Final Project- ESG Sentiment Analysis\Fashion Retail
Companies: 15

Companies to analyze:
  - Aritzia (Low controversy)
  - Fast Retailing (Medium controversy)
  - GAP Inc (Medium controversy)
  - H&M (High controversy)
  - Inditex (Medium controversy)
  - Levi's (Low controversy)
  - LVMH (Low controversy)
  - Lululemon (Low controversy)
  - Marks & Spencer (Low controversy)
  - Patagonia (Low controversy)
  - Prada (Low controversy)
  - Ralph Lauren (Low controversy)
  - Shein (High controversy)
  - TJX Companies (Medium controversy)
  - Victoria's Secret (Medium controversy)

############################################################
INDUSTRY: Fashion Retail
############################################################

============================================================
Analyzing: Aritzia
Industry: Fashion Retail | Controversy: Low
============================================================
Extracted 203,002 characters from 74 pages
Word count: 31,119

Sentiment:
  VADER: 1.000
  TextBlob: 0.087
  Positive sentences: 21.0%

Greenwashing:
  Risk: High
  Ratio: 0.40
  Aspirational words: 491
  Substantive words: 1243

ESG Focus:
  Environmental: 29.4%
  Social: 44.2%
  Governance: 26.4%
  Dominant: Social

Analysis complete for Aritzia


============================================================
Analyzing: Fast Retailing
Industry: Fashion Retail | Controversy: Medium
============================================================
Extracted 176,460 characters from 92 pages
Word count: 27,629

Sentiment:
  VADER: 1.000
  TextBlob: 0.147
  Positive sentences: 31.1%
```

```
Greenwashing:
  Risk: High
  Ratio: 0.63
  Aspirational words: 485
  Substantive words: 774

ESG Focus:
  Environmental: 27.6%
  Social: 44.4%
  Governance: 28.1%
  Dominant: Social

Analysis complete for Fast Retailing


============================================================
Analyzing: GAP Inc
Industry: Fashion Retail | Controversy: Medium
============================================================
Extracted 198,195 characters from 58 pages
Word count: 29,827

Sentiment:
  VADER: 1.000
  TextBlob: 0.099
  Positive sentences: 30.5%

Greenwashing:
  Risk: Medium
  Ratio: 0.27
  Aspirational words: 375
  Substantive words: 1365

ESG Focus:
  Environmental: 24.4%
  Social: 46.5%
  Governance: 29.1%
  Dominant: Social

Analysis complete for GAP Inc


============================================================
Analyzing: H&M
Industry: Fashion Retail | Controversy: High
============================================================
Extracted 654,300 characters from 87 pages
Word count: 105,539

Sentiment:
  VADER: 1.000
  TextBlob: 0.093
  Positive sentences: 27.2%

Greenwashing:
  Risk: Medium
  Ratio: 0.37
  Aspirational words: 1261
  Substantive words: 3369

ESG Focus:
  Environmental: 29.6%
  Social: 39.9%
  Governance: 30.5%
  Dominant: Social
```

```
Analysis complete for H&M


============================================================
Analyzing: Inditex
Industry: Fashion Retail | Controversy: Medium
============================================================
Extracted 945,159 characters from 296 pages
Word count: 146,665

Sentiment:
  VADER: 1.000
  TextBlob: 0.088
  Positive sentences: 27.9%

Greenwashing:
  Risk: Medium
  Ratio: 0.33
  Aspirational words: 1581
  Substantive words: 4796

ESG Focus:
  Environmental: 29.7%
  Social: 41.6%
  Governance: 28.7%
  Dominant: Social

Analysis complete for Inditex


============================================================
Analyzing: Levi's
Industry: Fashion Retail | Controversy: Low
============================================================
Extracted 83,720 characters from 25 pages
Word count: 14,344

Sentiment:
  VADER: 1.000
  TextBlob: 0.041
  Positive sentences: 4.2%

Greenwashing:
  Risk: Medium
  Ratio: 0.28
  Aspirational words: 176
  Substantive words: 628

ESG Focus:
  Environmental: 48.2%
  Social: 28.0%
  Governance: 23.7%
  Dominant: Environmental

Analysis complete for Levi's


============================================================
Analyzing: LVMH
Industry: Fashion Retail | Controversy: Low
============================================================
Extracted 341,595 characters from 162 pages
Word count: 54,192

Sentiment:
```

```
  VADER: 1.000
  TextBlob: 0.122
  Positive sentences: 35.0%

Greenwashing:
  Risk: High
  Ratio: 0.61
  Aspirational words: 936
  Substantive words: 1541

ESG Focus:
  Environmental: 31.5%
  Social: 51.4%
  Governance: 17.1%
  Dominant: Social

Analysis complete for LVMH


============================================================
Analyzing: Lululemon
Industry: Fashion Retail | Controversy: Low
============================================================
Extracted 316,886 characters from 92 pages
Word count: 66,347

Sentiment:
  VADER: 1.000
  TextBlob: 0.163
  Positive sentences: 2.8%

Greenwashing:
  Risk: Medium
  Ratio: 0.86
  Aspirational words: 50
  Substantive words: 58

ESG Focus:
  Environmental: 8.4%
  Social: 58.1%
  Governance: 33.5%
  Dominant: Social

Analysis complete for Lululemon


============================================================
Analyzing: Marks & Spencer
Industry: Fashion Retail | Controversy: Low
============================================================
Extracted 175,292 characters from 69 pages
Word count: 27,735

Sentiment:
  VADER: 1.000
  TextBlob: 0.132
  Positive sentences: 38.2%

Greenwashing:
  Risk: Medium
  Ratio: 0.33
  Aspirational words: 403
  Substantive words: 1213

ESG Focus:
  Environmental: 40.0%
```

```
  Social: 39.5%
  Governance: 20.5%
  Dominant: Environmental

Analysis complete for Marks & Spencer


============================================================
Analyzing: Patagonia
Industry: Fashion Retail | Controversy: Low
============================================================
Extracted 25,950 characters from 25 pages
Word count: 4,126

Sentiment:
  VADER: 1.000
  TextBlob: 0.176
  Positive sentences: 49.5%

Greenwashing:
  Risk: High
  Ratio: 0.37
  Aspirational words: 62
  Substantive words: 167

ESG Focus:
  Environmental: 27.9%
  Social: 55.6%
  Governance: 16.5%
  Dominant: Social

Analysis complete for Patagonia


============================================================
Analyzing: Prada
Industry: Fashion Retail | Controversy: Low
============================================================
Extracted 293,531 characters from 179 pages
Word count: 45,241

Sentiment:
  VADER: 1.000
  TextBlob: 0.087
  Positive sentences: 29.6%

Greenwashing:
  Risk: High
  Ratio: 0.59
  Aspirational words: 786
  Substantive words: 1342

ESG Focus:
  Environmental: 34.7%
  Social: 40.5%
  Governance: 24.8%
  Dominant: Social

Analysis complete for Prada


============================================================
Analyzing: Ralph Lauren
Industry: Fashion Retail | Controversy: Low
============================================================
Extracted 147,208 characters from 52 pages
```

```
Word count: 21,836

Sentiment:
  VADER: 1.000
  TextBlob: 0.065
  Positive sentences: 26.1%

Greenwashing:
  Risk: Medium
  Ratio: 0.26
  Aspirational words: 286
  Substantive words: 1081

ESG Focus:
  Environmental: 44.7%
  Social: 34.2%
  Governance: 21.1%
  Dominant: Environmental

Analysis complete for Ralph Lauren


============================================================
Analyzing: Shein
Industry: Fashion Retail | Controversy: High
============================================================
Extracted 369,529 characters from 111 pages
Word count: 55,544

Sentiment:
  VADER: 1.000
  TextBlob: 0.090
  Positive sentences: 32.0%

Greenwashing:
  Risk: High
  Ratio: 0.45
  Aspirational words: 851
  Substantive words: 1890

ESG Focus:
  Environmental: 34.3%
  Social: 41.5%
  Governance: 24.2%
  Dominant: Social

Analysis complete for Shein


============================================================
Analyzing: TJX Companies
Industry: Fashion Retail | Controversy: Medium
============================================================
Extracted 231,777 characters from 82 pages
Word count: 36,409

Sentiment:
  VADER: 1.000
  TextBlob: 0.125
  Positive sentences: 39.2%

Greenwashing:
  Risk: High
  Ratio: 0.75
  Aspirational words: 723
  Substantive words: 969
```

```
ESG Focus:
  Environmental: 26.9%
  Social: 48.7%
  Governance: 24.4%
  Dominant: Social

Analysis complete for TJX Companies


============================================================
Analyzing: Victoria's Secret
Industry: Fashion Retail | Controversy: Medium
============================================================
Extracted 302,895 characters from 111 pages
Word count: 45,306

Sentiment:
  VADER: 1.000
  TextBlob: 0.124
  Positive sentences: 36.1%

Greenwashing:
  Risk: High
  Ratio: 0.62
  Aspirational words: 809
  Substantive words: 1308

ESG Focus:
  Environmental: 28.6%
  Social: 41.4%
  Governance: 30.0%
  Dominant: Social

Analysis complete for Victoria's Secret


Running TF-IDF analysis across 15 companies...

============================================================
Industry analysis complete: Fashion Retail
Companies analyzed: 15
============================================================


============================================================
RESULTS PREVIEW
============================================================
```

	company_name	industry	controversy_level	num_pages	word_count	text_content	vader_compo
0	Aritzia	Fashion Retail	Low	74	31119	FY2024 Aritzia Community ESG Report1 FY2024 Ar...	
1	Fast Retailing	Fashion Retail	Medium	92	27629	INTEGRATED REPORT 2024LifeWear, Changing the W...	
2	GAP Inc	Fashion Retail	Medium	58	29827	This report covers Gap Inc. s global operation...	
3	H&M	Fashion Retail	High	87	105539	H M GROUP ANNUAL SUST AINABILITY REPORT2024 ...	
4	Inditex	Fashion Retail	Medium	296	146665	Consolidated Statement of Non Financial Inform...	

5 rows × 34 columns

Key Metrics:

	company_name	controversy_level	vader_compound	risk_level	greenwashing_ratio	dominant_pi
0	Aritzia	Low	1.0	High	0.40	So
1	Fast Retailing	Medium	1.0	High	0.63	So
2	GAP Inc	Medium	1.0	Medium	0.27	So
3	H&M	High	1.0	Medium	0.37	So
4	Inditex	Medium	1.0	Medium	0.33	So
5	Levi's	Low	1.0	Medium	0.28	Environme
6	LVMH	Low	1.0	High	0.61	So
7	Lululemon	Low	1.0	Medium	0.86	So
8	Marks & Spencer	Low	1.0	Medium	0.33	Environme
9	Patagonia	Low	1.0	High	0.37	So
10	Prada	Low	1.0	High	0.59	So
11	Ralph Lauren	Low	1.0	Medium	0.26	Environme
12	Shein	High	1.0	High	0.45	So
13	TJX Companies	Medium	1.0	High	0.75	So
14	Victoria's Secret	Medium	1.0	High	0.62	So

```
Saving to Excel: Fashion_Retail_ESG_Analysis_20251012_1205.xlsx
✓ Saved to: Fashion_Retail_ESG_Analysis_20251012_1205.xlsx

============================================================
ANALYZING: Healthcare
============================================================
PDF folder: C:\Users\sonali\OneDrive\Desktop\Trimester I\AN6002 Analytics and ML in Busine
ss\Final Project- ESG Sentiment Analysis\Healthcare
Companies: 10

Companies to analyze:
  - Cardinal Health (Medium controversy)
  - Cencora (Low controversy)
  - Centene (Medium controversy)
  - The Cigna Group (Low controversy)
  - CVS Health (Medium controversy)
  - Elevance Health (Low controversy)
  - Humana (Low controversy)
  - McKesson (Low controversy)
  - UnitedHealth Group (Medium controversy)
  - Walgreens Boots Alliance (Medium controversy)

############################################################
INDUSTRY: Healthcare
############################################################

============================================================
Analyzing: Cardinal Health
Industry: Healthcare | Controversy: Medium
============================================================
Extracted 358,701 characters from 107 pages
Word count: 53,915

Sentiment:
  VADER: 1.000
  TextBlob: 0.085
  Positive sentences: 28.6%

Greenwashing:
  Risk: Medium
  Ratio: 0.31
  Aspirational words: 622
  Substantive words: 1999

ESG Focus:
  Environmental: 21.4%
  Social: 46.2%
  Governance: 32.4%
  Dominant: Social

Analysis complete for Cardinal Health


============================================================
Analyzing: Cencora
Industry: Healthcare | Controversy: Low
============================================================
Extracted 25,623 characters from 22 pages
Word count: 3,874

Sentiment:
  VADER: 1.000
  TextBlob: 0.141
  Positive sentences: 39.4%

Greenwashing:
```

```
  Risk: High
  Ratio: 0.64
  Aspirational words: 68
  Substantive words: 107

ESG Focus:
  Environmental: 20.8%
  Social: 56.9%
  Governance: 22.4%
  Dominant: Social

Analysis complete for Cencora


============================================================
Analyzing: Centene
Industry: Healthcare | Controversy: Medium
============================================================
Extracted 160,681 characters from 78 pages
Word count: 23,357

Sentiment:
  VADER: 1.000
  TextBlob: 0.160
  Positive sentences: 45.6%

Greenwashing:
  Risk: High
  Ratio: 0.45
  Aspirational words: 387
  Substantive words: 855

ESG Focus:
  Environmental: 12.8%
  Social: 61.9%
  Governance: 25.3%
  Dominant: Social

Analysis complete for Centene


============================================================
Analyzing: The Cigna Group
Industry: Healthcare | Controversy: Low
============================================================
Extracted 310,320 characters from 100 pages
Word count: 45,824

Sentiment:
  VADER: 1.000
  TextBlob: 0.089
  Positive sentences: 29.9%

Greenwashing:
  Risk: High
  Ratio: 0.35
  Aspirational words: 708
  Substantive words: 2036

ESG Focus:
  Environmental: 15.9%
  Social: 58.1%
  Governance: 26.0%
  Dominant: Social

Analysis complete for The Cigna Group
```

```
============================================================
Analyzing: CVS Health
Industry: Healthcare | Controversy: Medium
============================================================
Extracted 189,927 characters from 98 pages
Word count: 29,568

Sentiment:
  VADER: 1.000
  TextBlob: 0.096
  Positive sentences: 26.1%

Greenwashing:
  Risk: High
  Ratio: 0.35
  Aspirational words: 456
  Substantive words: 1292

ESG Focus:
  Environmental: 26.8%
  Social: 41.8%
  Governance: 31.4%
  Dominant: Social

Analysis complete for CVS Health


============================================================
Analyzing: Elevance Health
Industry: Healthcare | Controversy: Low
============================================================
Extracted 88,147 characters from 34 pages
Word count: 13,165

Sentiment:
  VADER: 1.000
  TextBlob: 0.117
  Positive sentences: 30.8%

Greenwashing:
  Risk: Medium
  Ratio: 0.35
  Aspirational words: 175
  Substantive words: 499

ESG Focus:
  Environmental: 20.1%
  Social: 50.0%
  Governance: 29.8%
  Dominant: Social

Analysis complete for Elevance Health


============================================================
Analyzing: Humana
Industry: Healthcare | Controversy: Low
============================================================
Extracted 512,948 characters from 136 pages
Word count: 78,714

Sentiment:
  VADER: 1.000
  TextBlob: 0.093
```

```
  Positive sentences: 27.8%

Greenwashing:
  Risk: Medium
  Ratio: 0.20
  Aspirational words: 667
  Substantive words: 3278

ESG Focus:
  Environmental: 22.2%
  Social: 42.0%
  Governance: 35.8%
  Dominant: Social

Analysis complete for Humana


============================================================
Analyzing: McKesson
Industry: Healthcare | Controversy: Low
============================================================
Extracted 150,395 characters from 71 pages
Word count: 23,185

Sentiment:
  VADER: 1.000
  TextBlob: 0.127
  Positive sentences: 34.2%

Greenwashing:
  Risk: High
  Ratio: 0.50
  Aspirational words: 368
  Substantive words: 743

ESG Focus:
  Environmental: 24.9%
  Social: 49.1%
  Governance: 26.0%
  Dominant: Social

Analysis complete for McKesson


============================================================
Analyzing: UnitedHealth Group
Industry: Healthcare | Controversy: Medium
============================================================
Extracted 186,626 characters from 87 pages
Word count: 28,646

Sentiment:
  VADER: 1.000
  TextBlob: 0.156
  Positive sentences: 39.4%

Greenwashing:
  Risk: High
  Ratio: 0.35
  Aspirational words: 497
  Substantive words: 1436

ESG Focus:
  Environmental: 20.4%
  Social: 55.0%
  Governance: 24.6%
```

```
  Dominant: Social

Analysis complete for UnitedHealth Group


============================================================
Analyzing: Walgreens Boots Alliance
Industry: Healthcare | Controversy: Medium
============================================================
Extracted 214,414 characters from 82 pages
Word count: 32,781

Sentiment:
  VADER: 1.000
  TextBlob: 0.159
  Positive sentences: 42.3%

Greenwashing:
  Risk: High
  Ratio: 0.74
  Aspirational words: 681
  Substantive words: 915

ESG Focus:
  Environmental: 29.0%
  Social: 57.8%
  Governance: 13.2%
  Dominant: Social

Analysis complete for Walgreens Boots Alliance


Running TF-IDF analysis across 10 companies...

============================================================
Industry analysis complete: Healthcare
Companies analyzed: 10
============================================================


============================================================
RESULTS PREVIEW
============================================================
```

	company_name	industry	controversy_level	num_pages	word_count	text_content	vader_co
0	Cardinal Health	Healthcare	Medium	107	53915	Published February , Environmental, social an...	
1	Cencora	Healthcare	Low	22	3874	Corporate Responsibility Summary Report 2Tabl...	
2	Centene	Healthcare	Medium	78	23357	CORPORATE RESPONSIBILITY REPORT Empowering Hea...	
3	The Cigna Group	Healthcare	Low	100	45824	Purpose Performance CORPORATE IMPACT REPORT...	
4	CVS Health	Healthcare	Medium	98	29568	Healthy Impact Report Appendix Table of Cont...	

5 rows × 34 columns

Key Metrics:

	company_name	controversy_level	vader_compound	risk_level	greenwashing_ratio	dominant_pill
0	Cardinal Health	Medium	1.0000	Medium	0.31	Soc
1	Cencora	Low	0.9999	High	0.64	Soc
2	Centene	Medium	1.0000	High	0.45	Soc
3	The Cigna Group	Low	1.0000	High	0.35	Soc
4	CVS Health	Medium	1.0000	High	0.35	Soc
5	Elevance Health	Low	1.0000	Medium	0.35	Soc
6	Humana	Low	1.0000	Medium	0.20	Soc
7	McKesson	Low	1.0000	High	0.50	Soc
8	UnitedHealth Group	Medium	1.0000	High	0.35	Soc
9	Walgreens Boots Alliance	Medium	1.0000	High	0.74	Soc

```
Saving to Excel: Healthcare_ESG_Analysis_20251012_1213.xlsx
✓ Saved to: Healthcare_ESG_Analysis_20251012_1213.xlsx

============================================================
ANALYZING: Oil & Gas
============================================================
PDF folder: C:\Users\sonali\OneDrive\Desktop\Trimester I\AN6002 Analytics and ML in Busine
ss\Final Project- ESG Sentiment Analysis\Oil & Gas
Companies: 9

Companies to analyze:
  - Chevron (High controversy)
  - Enbridge (Medium controversy)
  - ExxonMobil (High controversy)
  - Indian Oil (High controversy)
  - Marathon Petroleum (High controversy)
  - Occidental Petroleum (High controversy)
  - Saudi Aramco (High controversy)
  - Shell (High controversy)
  - Southern Company (Medium controversy)

############################################################
INDUSTRY: Oil & Gas
############################################################

============================================================
Analyzing: Chevron
Industry: Oil & Gas | Controversy: High
============================================================
Extracted 74,798 characters from 32 pages
Word count: 10,787

Sentiment:
  VADER: 1.000
  TextBlob: 0.038
  Positive sentences: 9.7%

Greenwashing:
  Risk: Medium
  Ratio: 0.16
  Aspirational words: 115
  Substantive words: 732

ESG Focus:
  Environmental: 48.5%
  Social: 30.4%
  Governance: 21.1%
  Dominant: Environmental

Analysis complete for Chevron


============================================================
Analyzing: Enbridge
Industry: Oil & Gas | Controversy: Medium
============================================================
Extracted 462,716 characters from 116 pages
Word count: 70,892

Sentiment:
  VADER: 1.000
  TextBlob: 0.101
  Positive sentences: 31.4%

Greenwashing:
  Risk: High
```

```
  Ratio: 0.57
  Aspirational words: 1253
  Substantive words: 2193

ESG Focus:
  Environmental: 39.3%
  Social: 37.3%
  Governance: 23.4%
  Dominant: Environmental

Analysis complete for Enbridge


============================================================
Analyzing: ExxonMobil
Industry: Oil & Gas | Controversy: High
============================================================
Extracted 228,720 characters from 72 pages
Word count: 35,992

Sentiment:
  VADER: 1.000
  TextBlob: 0.101
  Positive sentences: 17.8%

Greenwashing:
  Risk: High
  Ratio: 0.52
  Aspirational words: 664
  Substantive words: 1284

ESG Focus:
  Environmental: 34.5%
  Social: 45.4%
  Governance: 20.1%
  Dominant: Social

Analysis complete for ExxonMobil


============================================================
Analyzing: Indian Oil
Industry: Oil & Gas | Controversy: High
============================================================
Extracted 301,289 characters from 148 pages
Word count: 41,958

Sentiment:
  VADER: 1.000
  TextBlob: 0.071
  Positive sentences: 22.2%

Greenwashing:
  Risk: High
  Ratio: 0.40
  Aspirational words: 708
  Substantive words: 1765

ESG Focus:
  Environmental: 39.6%
  Social: 38.6%
  Governance: 21.8%
  Dominant: Environmental

Analysis complete for Indian Oil
```

```
============================================================
Analyzing: Marathon Petroleum
Industry: Oil & Gas | Controversy: High
============================================================
Extracted 263,366 characters from 59 pages
Word count: 38,823

Sentiment:
  VADER: 1.000
  TextBlob: 0.107
  Positive sentences: 22.2%

Greenwashing:
  Risk: High
  Ratio: 0.40
  Aspirational words: 584
  Substantive words: 1458

ESG Focus:
  Environmental: 30.7%
  Social: 45.3%
  Governance: 24.0%
  Dominant: Social

Analysis complete for Marathon Petroleum


============================================================
Analyzing: Occidental Petroleum
Industry: Oil & Gas | Controversy: High
============================================================
Extracted 248,849 characters from 91 pages
Word count: 39,613

Sentiment:
  VADER: 1.000
  TextBlob: 0.080
  Positive sentences: 21.7%

Greenwashing:
  Risk: High
  Ratio: 0.40
  Aspirational words: 642
  Substantive words: 1604

ESG Focus:
  Environmental: 42.4%
  Social: 34.7%
  Governance: 22.9%
  Dominant: Environmental

Analysis complete for Occidental Petroleum


============================================================
Analyzing: Saudi Aramco
Industry: Oil & Gas | Controversy: High
============================================================
Extracted 383,158 characters from 71 pages
Word count: 58,380

Sentiment:
  VADER: 1.000
  TextBlob: 0.081
  Positive sentences: 22.3%
```

```
Greenwashing:
  Risk: High
  Ratio: 0.40
  Aspirational words: 925
  Substantive words: 2317

ESG Focus:
  Environmental: 51.8%
  Social: 32.3%
  Governance: 15.9%
  Dominant: Environmental

Analysis complete for Saudi Aramco


============================================================
Analyzing: Shell
Industry: Oil & Gas | Controversy: High
============================================================
Extracted 331,913 characters from 98 pages
Word count: 69,073

Sentiment:
  VADER: 1.000
  TextBlob: 0.063
  Positive sentences: 9.7%

Greenwashing:
  Risk: Low
  Ratio: 0.27
  Aspirational words: 313
  Substantive words: 1166

ESG Focus:
  Environmental: 47.4%
  Social: 42.8%
  Governance: 9.8%
  Dominant: Environmental

Analysis complete for Shell


============================================================
Analyzing: Southern Company
Industry: Oil & Gas | Controversy: Medium
============================================================
Extracted 100,279 characters from 29 pages
Word count: 14,919

Sentiment:
  VADER: 1.000
  TextBlob: 0.112
  Positive sentences: 40.2%

Greenwashing:
  Risk: High
  Ratio: 0.81
  Aspirational words: 335
  Substantive words: 414

ESG Focus:
  Environmental: 39.8%
  Social: 37.7%
  Governance: 22.5%
  Dominant: Environmental
```

```
Analysis complete for Southern Company


Running TF-IDF analysis across 9 companies...

============================================================
Industry analysis complete: Oil & Gas
Companies analyzed: 9
============================================================


============================================================
RESULTS PREVIEW
============================================================
```

	company_name	industry	controversy_level	num_pages	word_count	text_content	vade
0	Chevron	Oil & Gas	High	32	10787	corporate sustainability highlights Chevron ...	
1	Enbridge	Oil & Gas	Medium	116	70892	Sustainability Report T able of contents In t...	
2	ExxonMobil	Oil & Gas	High	72	35992	Our view to Sustainability Report April Exx...	
3	Indian Oil	Oil & Gas	High	148	41958	SUSTAINABILITY REPORT About the ReportEnviro...	
4	Marathon Petroleum	Oil & Gas	High	59	38823	SUSTAINABILITYDriven S U S T A I N A B I L...	

5 rows × 34 columns

```
Key Metrics:
```

	company_name	controversy_level	vader_compound	risk_level	greenwashing_ratio	dominant_pill
0	Chevron	High	0.9997	Medium	0.16	Environment
1	Enbridge	Medium	1.0000	High	0.57	Environment
2	ExxonMobil	High	1.0000	High	0.52	Soc
3	Indian Oil	High	1.0000	High	0.40	Environment
4	Marathon Petroleum	High	1.0000	High	0.40	Soc
5	Occidental Petroleum	High	1.0000	High	0.40	Environment
6	Saudi Aramco	High	1.0000	High	0.40	Environment
7	Shell	High	0.9999	Low	0.27	Environment
8	Southern Company	Medium	1.0000	High	0.81	Environment

```
Saving to Excel: Oil_&_Gas_ESG_Analysis_20251012_1225.xlsx
✓ Saved to: Oil_&_Gas_ESG_Analysis_20251012_1225.xlsx

============================================================
ANALYZING: Agriculture
============================================================
PDF folder: C:\Users\sonali\OneDrive\Desktop\Trimester I\AN6002 Analytics and ML in Busine
ss\Final Project- ESG Sentiment Analysis\Agriculture
Companies: 11

Companies to analyze:
  - Astra Agro Lestari (High controversy)
  - Cargill (High controversy)
  - COFCO (Medium controversy)
  - DCM Shriram (Medium controversy)
  - Golden Agri-Resources (High controversy)
  - GrainCorp (Low controversy)
  - Kuala Lumpur Kepong (High controversy)
  - KWS (Low controversy)
  - M.P. Evans (High controversy)
  - Sakata (Low controversy)
  - Tessenderlo Group (Medium controversy)

############################################################
INDUSTRY: Agriculture
############################################################

============================================================
Analyzing: Astra Agro Lestari
Industry: Agriculture | Controversy: High
============================================================
Extracted 839,709 characters from 245 pages
Word count: 115,173

Sentiment:
  VADER: 1.000
  TextBlob: 0.083
  Positive sentences: 13.2%

Greenwashing:
  Risk: Medium
  Ratio: 0.36
  Aspirational words: 1079
  Substantive words: 3010

ESG Focus:
  Environmental: 29.6%
  Social: 43.9%
  Governance: 26.5%
  Dominant: Social

Analysis complete for Astra Agro Lestari


============================================================
Analyzing: Cargill
Industry: Agriculture | Controversy: High
============================================================
Extracted 286,248 characters from 122 pages
Word count: 44,556

Sentiment:
  VADER: 1.000
  TextBlob: 0.126
  Positive sentences: 36.7%
```

```
Greenwashing:
  Risk: Medium
  Ratio: 0.38
  Aspirational words: 661
  Substantive words: 1732

ESG Focus:
  Environmental: 41.1%
  Social: 40.5%
  Governance: 18.4%
  Dominant: Environmental

Analysis complete for Cargill


============================================================
Analyzing: COFCO
Industry: Agriculture | Controversy: Medium
============================================================
Extracted 391,417 characters from 114 pages
Word count: 60,357

Sentiment:
  VADER: 1.000
  TextBlob: 0.128
  Positive sentences: 39.1%

Greenwashing:
  Risk: High
  Ratio: 0.44
  Aspirational words: 931
  Substantive words: 2102

ESG Focus:
  Environmental: 41.2%
  Social: 39.2%
  Governance: 19.5%
  Dominant: Environmental

Analysis complete for COFCO


============================================================
Analyzing: DCM Shriram
Industry: Agriculture | Controversy: Medium
============================================================
Extracted 294,210 characters from 94 pages
Word count: 43,162

Sentiment:
  VADER: 1.000
  TextBlob: 0.087
  Positive sentences: 25.1%

Greenwashing:
  Risk: High
  Ratio: 0.50
  Aspirational words: 742
  Substantive words: 1496

ESG Focus:
  Environmental: 31.6%
  Social: 40.0%
  Governance: 28.4%
  Dominant: Social
```

```
Analysis complete for DCM Shriram


============================================================
Analyzing: Golden Agri-Resources
Industry: Agriculture | Controversy: High
============================================================
Extracted 317,514 characters from 131 pages
Word count: 46,674

Sentiment:
  VADER: 1.000
  TextBlob: 0.125
  Positive sentences: 33.5%

Greenwashing:
  Risk: High
  Ratio: 0.41
  Aspirational words: 761
  Substantive words: 1834

ESG Focus:
  Environmental: 41.6%
  Social: 35.2%
  Governance: 23.2%
  Dominant: Environmental

Analysis complete for Golden Agri-Resources


============================================================
Analyzing: GrainCorp
Industry: Agriculture | Controversy: Low
============================================================
Extracted 257,327 characters from 95 pages
Word count: 38,003

Sentiment:
  VADER: 1.000
  TextBlob: 0.106
  Positive sentences: 32.4%

Greenwashing:
  Risk: High
  Ratio: 0.51
  Aspirational words: 624
  Substantive words: 1232

ESG Focus:
  Environmental: 32.6%
  Social: 41.4%
  Governance: 26.0%
  Dominant: Social

Analysis complete for GrainCorp


============================================================
Analyzing: Kuala Lumpur Kepong
Industry: Agriculture | Controversy: High
============================================================
Extracted 255,474 characters from 91 pages
Word count: 38,066

Sentiment:
  VADER: 1.000
```

```
  TextBlob: 0.106
  Positive sentences: 30.6%

Greenwashing:
  Risk: Medium
  Ratio: 0.37
  Aspirational words: 523
  Substantive words: 1417

ESG Focus:
  Environmental: 26.7%
  Social: 36.5%
  Governance: 36.8%
  Dominant: Governance

Analysis complete for Kuala Lumpur Kepong


============================================================
Analyzing: KWS
Industry: Agriculture | Controversy: Low
============================================================
Error reading C:\Users\sonali\OneDrive\Desktop\Trimester I\AN6002 Analytics and ML in Busi
ness\Final Project- ESG Sentiment Analysis\Agriculture\KWS (DE).pdf: PyCryptodome is requi
red for AES algorithm
Failed to extract text

============================================================
Analyzing: M.P. Evans
Industry: Agriculture | Controversy: High
============================================================
Extracted 99,372 characters from 44 pages
Word count: 15,635

Sentiment:
  VADER: 1.000
  TextBlob: 0.057
  Positive sentences: 21.1%

Greenwashing:
  Risk: Medium
  Ratio: 0.33
  Aspirational words: 201
  Substantive words: 607

ESG Focus:
  Environmental: 49.8%
  Social: 21.0%
  Governance: 29.2%
  Dominant: Environmental

Analysis complete for M.P. Evans


============================================================
Analyzing: Sakata
Industry: Agriculture | Controversy: Low
============================================================
Extracted 161,679 characters from 23 pages
Word count: 25,466

Sentiment:
  VADER: 1.000
  TextBlob: 0.107
  Positive sentences: 33.3%
```

```
Greenwashing:
  Risk: High
  Ratio: 0.72
  Aspirational words: 445
  Substantive words: 622

ESG Focus:
  Environmental: 25.6%
  Social: 42.7%
  Governance: 31.7%
  Dominant: Social

Analysis complete for Sakata


============================================================
Analyzing: Tessenderlo Group
Industry: Agriculture | Controversy: Medium
============================================================
Extracted 174,812 characters from 83 pages
Word count: 27,695

Sentiment:
  VADER: 1.000
  TextBlob: 0.107
  Positive sentences: 28.3%

Greenwashing:
  Risk: Medium
  Ratio: 0.34
  Aspirational words: 378
  Substantive words: 1123

ESG Focus:
  Environmental: 43.6%
  Social: 35.6%
  Governance: 20.9%
  Dominant: Environmental

Analysis complete for Tessenderlo Group


Running TF-IDF analysis across 10 companies...

============================================================
Industry analysis complete: Agriculture
Companies analyzed: 10
============================================================


============================================================
RESULTS PREVIEW
============================================================
```

	company_name	industry	controversy_level	num_pages	word_count	text_content	vader_co
0	Astra Agro Lestari	Agriculture	High	245	115173	Sustainability Report PT Astra Agro Lestari ...	
1	Cargill	Agriculture	High	122	44556	Impact Report Table of contents Overview Lett...	
2	COFCO	Agriculture	Medium	114	60357	Sustainability Report COFCO International Ltd...	
3	DCM Shriram	Agriculture	Medium	94	43162	Rooted in Trust Ready for Tomorrow SUSTAINABIL...	
4	Golden Agri-Resources	Agriculture	High	131	46674	SUSTAINABILITY REPORT GOLDEN AGRI RESOURCES L...	

5 rows × 34 columns

Key Metrics:

	company_name	controversy_level	vader_compound	risk_level	greenwashing_ratio	dominant_pill
0	Astra Agro Lestari	High	1.0000	Medium	0.36	Soc
1	Cargill	High	1.0000	Medium	0.38	Environment
2	COFCO	Medium	1.0000	High	0.44	Environment
3	DCM Shriram	Medium	1.0000	High	0.50	Soc
4	Golden Agri-Resources	High	1.0000	High	0.41	Environment
5	GrainCorp	Low	1.0000	High	0.51	Soc
6	Kuala Lumpur Kepong	High	1.0000	Medium	0.37	Governan
7	M.P. Evans	High	0.9999	Medium	0.33	Environment
8	Sakata	Low	1.0000	High	0.72	Soc
9	Tessenderlo Group	Medium	1.0000	Medium	0.34	Environment

```
Saving to Excel: Agriculture_ESG_Analysis_20251012_1239.xlsx
✓ Saved to: Agriculture_ESG_Analysis_20251012_1239.xlsx

============================================================
ANALYZING: Pharma
============================================================
PDF folder: C:\Users\sonali\OneDrive\Desktop\Trimester I\AN6002 Analytics and ML in Busine
ss\Final Project- ESG Sentiment Analysis\Pharma
Companies: 10

Companies to analyze:
  - AbbVie (Medium controversy)
  - Agios Pharmaceuticals (Low controversy)
  - AstraZeneca (Low controversy)
  - Bristol Myers Squibb (Low controversy)
  - Daiichi Sankyo (Low controversy)
  - GSK (Medium controversy)
  - Merck (Low controversy)
  - Pfizer (Medium controversy)
  - Roche (Low controversy)
  - Takeda (Low controversy)

############################################################
INDUSTRY: Pharma
############################################################

============================================================
Analyzing: AbbVie
Industry: Pharma | Controversy: Medium
============================================================
Extracted 145,797 characters from 54 pages
Word count: 22,240

Sentiment:
  VADER: 1.000
  TextBlob: 0.111
  Positive sentences: 31.8%

Greenwashing:
  Risk: High
  Ratio: 0.56
  Aspirational words: 393
  Substantive words: 702

ESG Focus:
  Environmental: 19.3%
  Social: 52.2%
  Governance: 28.4%
  Dominant: Social

Analysis complete for AbbVie


============================================================
Analyzing: Agios Pharmaceuticals
Industry: Pharma | Controversy: Low
============================================================
Extracted 84,639 characters from 47 pages
Word count: 13,046

Sentiment:
  VADER: 1.000
  TextBlob: 0.166
  Positive sentences: 45.8%

Greenwashing:
```

```
  Risk: High
  Ratio: 1.11
  Aspirational words: 333
  Substantive words: 301

ESG Focus:
  Environmental: 11.6%
  Social: 64.1%
  Governance: 24.3%
  Dominant: Social

Analysis complete for Agios Pharmaceuticals


============================================================
Analyzing: AstraZeneca
Industry: Pharma | Controversy: Low
============================================================
Extracted 88,028 characters from 47 pages
Word count: 14,139

Sentiment:
  VADER: 1.000
  TextBlob: 0.132
  Positive sentences: 41.1%

Greenwashing:
  Risk: High
  Ratio: 0.92
  Aspirational words: 286
  Substantive words: 310

ESG Focus:
  Environmental: 40.0%
  Social: 51.9%
  Governance: 8.1%
  Dominant: Social

Analysis complete for AstraZeneca


============================================================
Analyzing: Bristol Myers Squibb
Industry: Pharma | Controversy: Low
============================================================
Extracted 337,032 characters from 118 pages
Word count: 50,035

Sentiment:
  VADER: 1.000
  TextBlob: 0.107
  Positive sentences: 35.8%

Greenwashing:
  Risk: High
  Ratio: 0.69
  Aspirational words: 1273
  Substantive words: 1845

ESG Focus:
  Environmental: 20.4%
  Social: 53.6%
  Governance: 26.1%
  Dominant: Social

Analysis complete for Bristol Myers Squibb
```

```
============================================================
Analyzing: Daiichi Sankyo
Industry: Pharma | Controversy: Low
============================================================
Extracted 107,149 characters from 20 pages
Word count: 16,054

Sentiment:
  VADER: 1.000
  TextBlob: 0.056
  Positive sentences: 17.6%

Greenwashing:
  Risk: High
  Ratio: 0.41
  Aspirational words: 252
  Substantive words: 617

ESG Focus:
  Environmental: 25.6%
  Social: 43.9%
  Governance: 30.5%
  Dominant: Social

Analysis complete for Daiichi Sankyo


============================================================
Analyzing: GSK
Industry: Pharma | Controversy: Medium
============================================================
Extracted 249,536 characters from 62 pages
Word count: 38,393

Sentiment:
  VADER: 1.000
  TextBlob: 0.097
  Positive sentences: 28.5%

Greenwashing:
  Risk: Medium
  Ratio: 0.26
  Aspirational words: 573
  Substantive words: 2241

ESG Focus:
  Environmental: 17.7%
  Social: 54.3%
  Governance: 28.0%
  Dominant: Social

Analysis complete for GSK


============================================================
Analyzing: Merck
Industry: Pharma | Controversy: Low
============================================================
Extracted 15,021 characters from 5 pages
Word count: 2,329

Sentiment:
  VADER: 0.999
  TextBlob: 0.151
```

```
  Positive sentences: 39.1%

Greenwashing:
  Risk: High
  Ratio: 0.37
  Aspirational words: 56
  Substantive words: 152

ESG Focus:
  Environmental: 33.3%
  Social: 42.6%
  Governance: 24.1%
  Dominant: Social

Analysis complete for Merck


============================================================
Analyzing: Pfizer
Industry: Pharma | Controversy: Medium
============================================================
Extracted 88,008 characters from 42 pages
Word count: 13,317

Sentiment:
  VADER: 1.000
  TextBlob: 0.071
  Positive sentences: 18.9%

Greenwashing:
  Risk: Medium
  Ratio: 0.16
  Aspirational words: 160
  Substantive words: 973

ESG Focus:
  Environmental: 27.2%
  Social: 33.5%
  Governance: 39.3%
  Dominant: Governance

Analysis complete for Pfizer


============================================================
Analyzing: Roche
Industry: Pharma | Controversy: Low
============================================================
Extracted 36,715 characters from 23 pages
Word count: 4,532

Sentiment:
  VADER: 1.000
  TextBlob: 0.081
  Positive sentences: 35.5%

Greenwashing:
  Risk: High
  Ratio: 0.66
  Aspirational words: 83
  Substantive words: 125

ESG Focus:
  Environmental: 57.5%
  Social: 36.6%
  Governance: 5.9%
```

```
  Dominant: Environmental

Analysis complete for Roche


============================================================
Analyzing: Takeda
Industry: Pharma | Controversy: Low
============================================================
Error reading C:\Users\sonali\OneDrive\Desktop\Trimester I\AN6002 Analytics and ML in Busi
ness\Final Project- ESG Sentiment Analysis\Pharma\Takeda.pdf: PyCryptodome is required for
AES algorithm
Failed to extract text

Running TF-IDF analysis across 9 companies...

============================================================
Industry analysis complete: Pharma
Companies analyzed: 9
============================================================


============================================================
RESULTS PREVIEW
============================================================
```

	company_name	industry	controversy_level	num_pages	word_count	text_content	vader_comp
0	AbbVie	Pharma	Medium	54	22240	ESG Action Report ESG Action Report Disclosur...	
1	Agios Pharmaceuticals	Pharma	Low	47	13046	Environmental, Social and Governance Report A...	
2	AstraZeneca	Pharma	Low	47	14139	Our Sustainability Impact Building a healthy f...	
3	Bristol Myers Squibb	Pharma	Low	118	50035	Building a Better Future Bristol Myers Squibb...	
4	Daiichi Sankyo	Pharma	Low	20	16054	ESG DataExternal ESG Evaluations Sustainabilit...	

5 rows × 34 columns

```
Key Metrics:
```

	company_name	controversy_level	vader_compound	risk_level	greenwashing_ratio	dominant_pill
0	AbbVie	Medium	1.0000	High	0.56	Soc
1	Agios Pharmaceuticals	Low	1.0000	High	1.11	Soc
2	AstraZeneca	Low	1.0000	High	0.92	Soc
3	Bristol Myers Squibb	Low	1.0000	High	0.69	Soc
4	Daiichi Sankyo	Low	1.0000	High	0.41	Soc
5	GSK	Medium	1.0000	Medium	0.26	Soc
6	Merck	Low	0.9991	High	0.37	Soc
7	Pfizer	Medium	0.9999	Medium	0.16	Governan
8	Roche	Low	0.9999	High	0.66	Environmen

```
Saving to Excel: Pharma_ESG_Analysis_20251012_1243.xlsx
✓ Saved to: Pharma_ESG_Analysis_20251012_1243.xlsx

======================================================================
ALL INDUSTRIES COMPLETE
======================================================================
✓ Fashion Retail: 15 companies analyzed
  Excel file: Fashion_Retail_ESG_Analysis_20251012_1205.xlsx
✓ Healthcare: 10 companies analyzed
  Excel file: Healthcare_ESG_Analysis_20251012_1213.xlsx
✓ Oil & Gas: 9 companies analyzed
  Excel file: Oil_&_Gas_ESG_Analysis_20251012_1225.xlsx
✓ Agriculture: 10 companies analyzed
  Excel file: Agriculture_ESG_Analysis_20251012_1239.xlsx
✓ Pharma: 9 companies analyzed
  Excel file: Pharma_ESG_Analysis_20251012_1243.xlsx

Creating master file: ALL_INDUSTRIES_ESG_Analysis_20251012_1243.xlsx
✓ Master file saved: ALL_INDUSTRIES_ESG_Analysis_20251012_1243.xlsx

======================================================================
ANALYSIS COMPLETE!
======================================================================
Total companies analyzed: 53
Individual industry files: 5
Combined master file: ALL_INDUSTRIES_ESG_Analysis_20251012_1243.xlsx
```