**Big data Project Report**

**A logo with text and a black background

Description automatically generated**

**MASTER OF BUSINESS ADMINISTRATION**

**(BUSINESS ANALYTICS)**

***Submitted by***

***Samrat Chakraborty***

***(Admission No: 23MB0057)***

***And***

***Anupurba Mukherjee***

***(Admission No: 23MB0009)***

***Under the supervision of***

***Prof. Shikha Singh***

***Professor***

**Decoding Honda’s Sustainability Narrative: A Topic Modeling and Industry 5.0 Benchmarking Study**

**1. Introduction**

As global companies increasingly rely on digital platforms to express their brand identity and strategic vision, analyzing the content of their official communications provides key insights into corporate priorities and positioning. In this project, we apply natural language processing (NLP) techniques—primarily web scraping and topic modeling—to uncover and interpret the central topics of **Honda Motor Company** as reflected in both its global website and its 2024 Sustainability Report.

The objective of this study is twofold:

* To identify and understand the **dominant communication topics** across Honda’s website and sustainability narratives.
* To assess how these topics align with **Industry 5.0 principles**, a modern industrial vision focusing on sustainability, human-centric innovation, technological resilience, and digital intelligence.

**2. Methodology**

**2.1 Web Scraping for Website Data**

We extracted textual content from Honda's global website using a custom web scraper. The scraping logic included not only the homepage but also all internal links, ensuring a comprehensive capture of Honda’s digital messaging. This included sections like:

* Investor Relations
* About Honda
* Sustainability and CSR
* Technology and Mobility Innovations
* Brand Philosophy

Text from these pages was cleaned and stored in a structured JSON format , making it suitable for further analysis.

**2.2 Preprocessing and Corpus Creation**

To prepare the text for analysis, we applied standard preprocessing techniques:

* Lowercasing
* Removal of stopwords and punctuation
* Tokenization into words
* Construction of a bag-of-words model using gensim

The resulting corpus allowed us to identify term frequencies and model topic distributions across all documents.

**3. Topic Modeling on Honda’s Website**

Using **Latent Dirichlet Allocation (LDA)**, we trained a model with 5 topics to explore the major topics in Honda's website content. Each topic is interpreted based on the most dominant keywords and supported by visualizations for clarity.

**3.1 Overall Word Cloud**

The word cloud revealed that **"honda"**, unsurprisingly, was the most dominant word across all web pages. Other frequently appearing terms included *mobility*, *dreams*, *corporate*, *power*, and *innovation*—indicating a strong brand-driven and forward-thinking communication style.

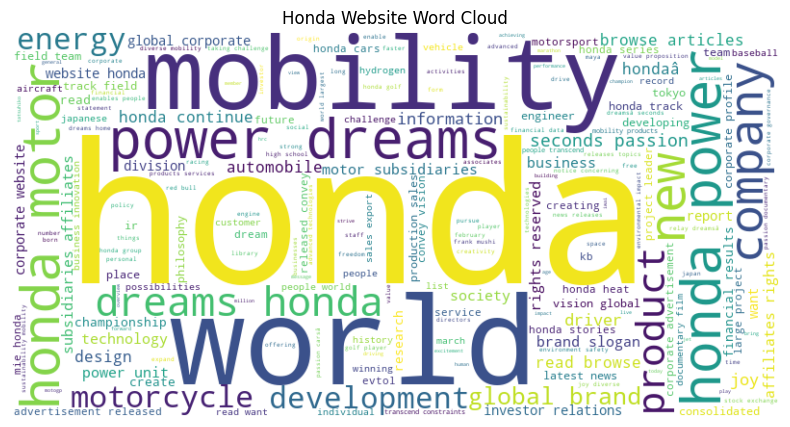


Figure 1: Honda Website Word Cloud Highlighting Dominant Brand Topics

**3.2 Identified Topics and Interpretation**

The five topics extracted from the LDA model represent distinct thematic pillars of Honda’s web communication.

|  |  |  |
| --- | --- | --- |
| **Topic** | **Focus Area** | **Interpretation** |
| Topic 0 | Brand Passion and Vision | Terms like *dreams*, *joy*, *mobility*, and *passion* emphasize Honda’s human-centric branding. It positions itself not just as a company, but as a dream enabler. |
| Topic 1 | Corporate Philosophy and Global Identity | With words like *corporate*, *philosophy*, *power*, and *world*, this topic reflects Honda’s self-perception as a global corporate citizen. |
| Topic 2 | Brand Rights and Legal Messaging | The frequent use of *rights*, *reserved*, *slogan*, and *march* shows Honda’s effort to protect and promote its brand identity legally. |
| Topic 3 | Racing Heritage and Technological Prowess | Words like *racing*, *championships*, and *technologies* signal the integration of speed, legacy, and innovation in Honda’s storytelling. |
| Topic 4 | Leadership and Business Structure | Terms such as *subsidiaries*, *team*, *investor*, and *leader* indicate the company’s openness about its business hierarchy and leadership. |

Table 1: Topics extracted from LDA model

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Figure 2: Topic-Specific Word Clouds and t-SNE Visualization of Website Topics  The t-SNE Inter-topic Distance Map visually illustrates how Honda’s website documents cluster around five distinct topics. Each point represents a document, colored by its dominant topic. The clear separation between clusters confirms that the topics—ranging from emotional branding and corporate identity to legal messaging, racing heritage, and business structure—are well-defined and non-overlapping. Notably, Topics 0 and 1 show slight proximity, reflecting shared elements around brand vision and corporate philosophy. The overall map validates the strength of the topic model and highlights the thematic clarity in Honda’s digital communication. |  |  |
|  |  |  |

**3.3 Document-Topic Distribution**

To evaluate how clearly the extracted topics map onto Honda’s website content, we examined the topic probabilities for individual documents. Some pages, such as those discussing Honda’s corporate philosophy and global presence, were almost entirely dominated by Topic 1. Others showed mixed associations—for instance, a page might reflect 40% content related to business structure (Topic 4) and 30% tied to brand vision (Topic 0). This distribution confirms that while some content is strongly aligned with specific topics, other pages incorporate multiple messaging dimensions, reflecting the multidimensional nature of Honda’s communication strategy.

**4. Topic Modeling on Honda’s Sustainability Report**To gain deeper insight into Honda’s environmental, social, and governance (ESG) communication strategy, we performed topic modeling on the 2024 Sustainability Report. The goal was to uncover key thematic areas in the company’s sustainability narrative and evaluate the coherence and clarity of these topics.

**4.1 Coherence Score Optimization**

To determine the ideal number of topics for modeling, we applied a coherence score evaluation across a range of topic numbers. Coherence score measures the semantic consistency of words within topics—higher scores indicate more interpretable topics.

As shown in the line chart below, the highest coherence score was achieved when the number of topics was set to three. Beyond this point, the scores declined significantly, suggesting that adding more topics would introduce noise or overlap rather than improving interpretability.

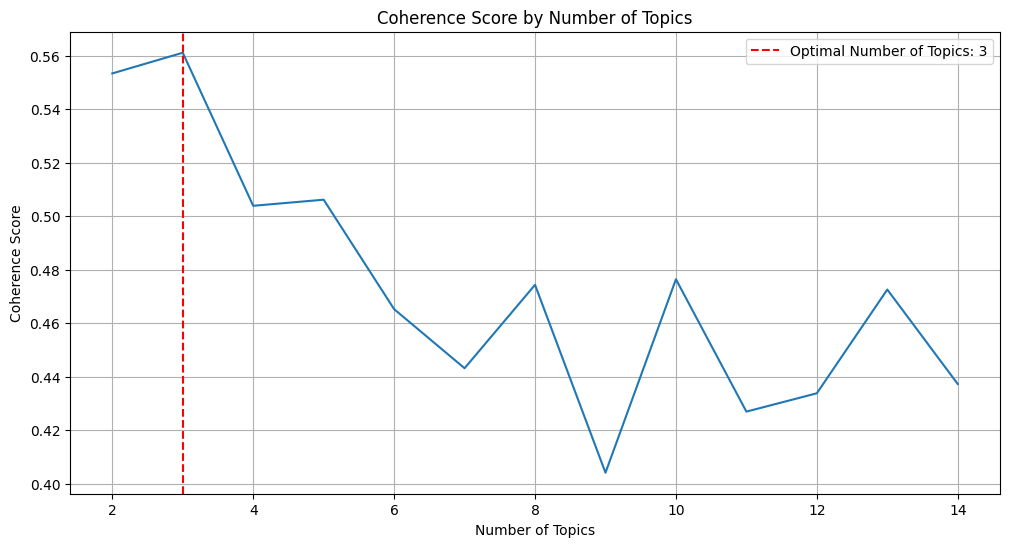
****

Figure 3: Coherence Score by Number of Topics for Sustainability Report

**4.2 Interpreting the Three Sustainability Topics**

The LDA model with three topics produced well-defined clusters, each highlighting a unique aspect of Honda’s sustainability communication. These topics reflect a balance between business responsibility, innovation, and legacy building.

|  |  |  |
| --- | --- | --- |
| **Topic** | **Thematic Focus** | **Key Interpretation** |
| **Topic 1** | Business Strategy & Governance | This topic highlights Honda’s focus on strategic planning, corporate leadership, and long-term value creation. Keywords such as *strategy*, *governance*, *value*, *theme*, and *management* indicate strong emphasis on internal governance structures and business ethics. |
| **Topic 2** | Innovation and Future Development | This topic reflects Honda’s forward-looking stance, with key terms like *technology*, *mobility*, *driving*, *dream*, and *development*. It illustrates the company’s commitment to innovation, environmental challenges, and future mobility solutions. |
| **Topic 3** | Legacy and Production Milestones | Comprising terms such as *launched*, *production*, *engine*, *motorcycle*, *japan*, and *automobile*, this topic celebrates Honda’s historical achievements, technological lineage, and manufacturing legacy. It positions Honda as a company with both rich heritage and modern relevance. |

Table 2: Sustainability Topics

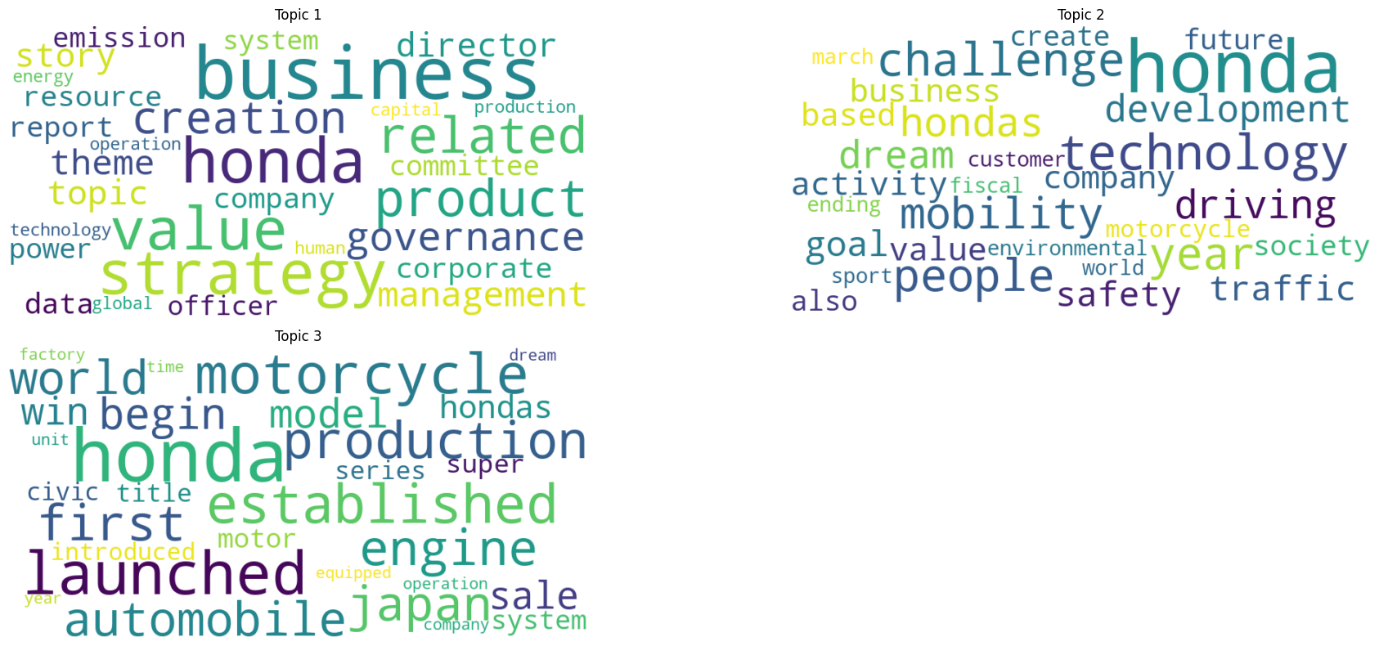


Figure 4: Word Clouds Depicting Key Topics from Honda’s Sustainability Report

**5. Similarity with Industry 5.0 Principles**

As part of this study, we assessed how closely Honda’s sustainability communication aligns with the evolving paradigm of **Industry 5.0**—a framework that emphasizes not only automation and technology, but also **human-centric innovation**, **resilience**, and **sustainability**. To do this, we compared the top keywords from each of the three identified topics with a curated list of over 80 keywords representing Industry 5.0 values and concepts. The comparison was conducted using **cosine similarity**, a common metric for measuring semantic overlap in text analytics.

**5.1 Quantitative Similarity Scores**

The results of the cosine similarity analysis are shown below. Topic 1, which focuses on business strategy and governance, had the **highest similarity score of 0.1378**, indicating modest alignment with Industry 5.0 topics—especially those related to corporate value creation and sustainable leadership. In contrast, Topics 2 and 3 had significantly lower similarity scores, suggesting limited overlap with Industry 5.0 ideals such as smart technologies, adaptive systems, and personalized manufacturing.

|  |  |
| --- | --- |
| **Topic** | **Cosine Similarity with Industry 5.0** |
| Topic 1 | 0.1378 |
| Topic 2 | 0.0372 |
| Topic 3 | 0.0277 |

Table 3: Quantitative Similarity Scores

**Average Alignment Score**: 0.0676

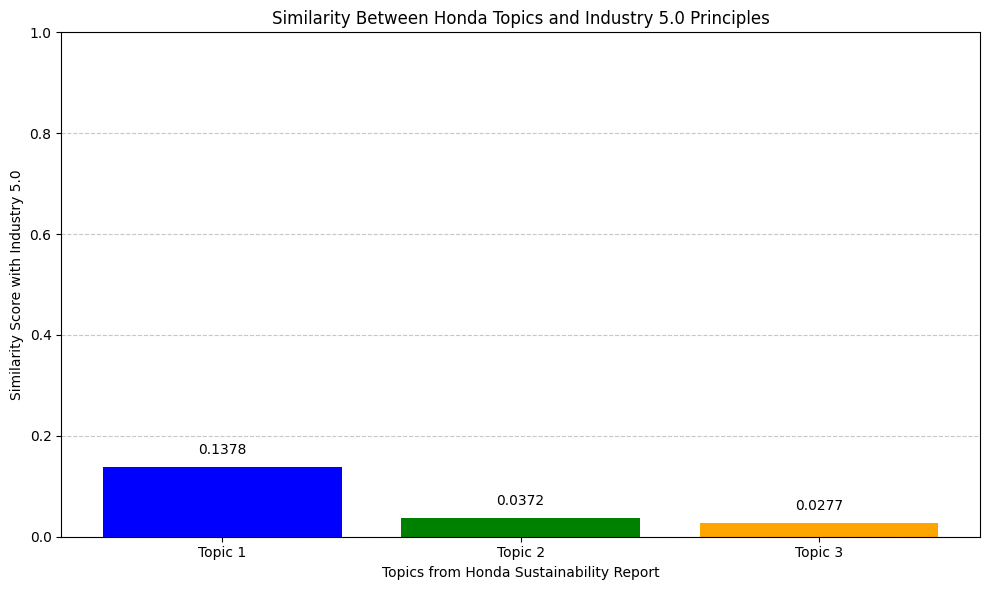


Figure 5: Similarity Between Honda Sustainability Topics and Industry 5.0 Principles

**5.2 Radar Chart Analysis by Industry 5.0 Dimensions**

To gain a more nuanced understanding of how Honda's sustainability communication aligns with the **core pillars of Industry 5.0**, we categorized the Industry 5.0 principles into five key dimensions: **Human-Centric Design**, **Sustainability**, **Technology**, **Innovation**, and **Resilience**. Each of the three identified topics from Honda’s sustainability report was then evaluated for its semantic similarity to these five categories using keyword-based analysis and cosine similarity measures.

The resulting **radar chart** visualizes the alignment of each topic across the five dimensions. **Topic 1**, which focuses on corporate governance and strategy, shows the strongest overall alignment, particularly in the areas of **Sustainability** and **Resilience**, suggesting that Honda’s communication reflects long-term value creation and responsibility. In contrast, **Topics 2 and 3**, which relate to innovation and production legacy respectively, display minimal alignment with the dimensions of **Technology** and **Human-Centric Design**. This indicates that while Honda’s messaging includes forward-looking elements like mobility and development, it currently lacks explicit references to **smart technologies**, **AI**, **worker empowerment**, or **personalized manufacturing**—hallmarks of Industry 5.0.

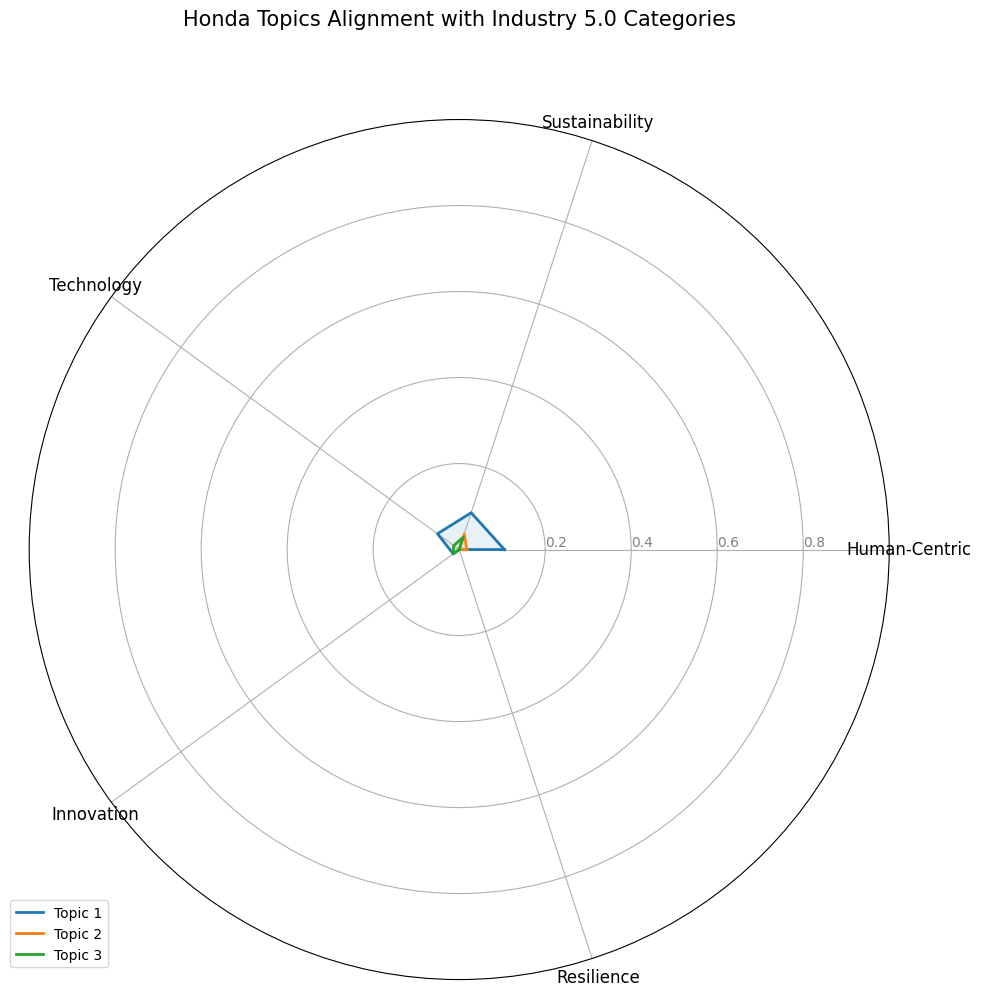


Figure 6: Radar Chart of Honda Topics Aligned with Industry 5.0 Dimensions

**Conclusion and Recommendations**

The study provides a comprehensive analysis of Honda's public communication through its global website and 2024 Sustainability Report, using topic modeling and semantic similarity techniques. The findings demonstrate that Honda is highly effective in articulating its **core values**, **corporate vision**, and **leadership strategies**, with consistent messaging around sustainability, governance, and innovation legacy.

However, while Honda's communication aligns well with traditional **ESG principles** and **Industry 4.0 frameworks**, it shows **limited alignment with the emerging standards of Industry 5.0**, which emphasize human-centric design, smart technologies, personalization, and socio-technical resilience.

**Key Strengths Identified**

* A clear and consistent brand philosophy built around dreams, passion, and mobility.
* Strong emphasis on sustainability, corporate governance, and long-term value creation.
* Balanced narrative that connects Honda’s legacy in manufacturing with its current innovation efforts.

**Observed Gaps**

* Limited references to **personalized manufacturing**, **adaptive systems**, or **human-machine collaboration**, which are central to Industry 5.0.
* Minimal alignment with **emerging technologies** such as **AI-driven automation**, **edge computing**, or **intelligent robotics**.
* Lack of storytelling around **worker empowerment**, **human-centric innovation**, or **collaborative technologies** in Honda’s sustainability discourse.

**Strategic Recommendations**

1. **Integrate Industry 5.0 Language and Topics**: Future communication—particularly in sustainability and innovation reporting—should explicitly reference concepts such as *customization*, *cyber-physical collaboration*, *intelligent automation*, and *worker-centered design*.
2. **Showcase Technological Foresight**: Honda should more prominently highlight its initiatives in **smart factories**, **digital twins**, **customizable electric vehicle (EV) platforms**, and **AI-powered safety systems**, which demonstrate alignment with Industry 5.0’s innovation focus.
3. **Humanize Technological Narratives**: Reports and brand messaging should increasingly reflect **human-AI collaboration**, **employee well-being**, and **resilience-building**, positioning Honda not just as a mobility innovator but as a human-centric, future-ready enterprise.