IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

# A TSP Framework for Verifying Cosmetic Sustainability Claims on ONDC

**Mr. P.J. Jambhulkar, Maitri More, Vaishnavi Rajput, Snehal Naik, Shruti Wawage**
Computer Engineering Department
SCTR's Pune Institute of Computer Technology, Pune, India
pjambhulkar@pict.edu, maitrimoreofficial@gmail.com
vaishnavirajput273@gmail.com, snehalnaik139@gmail.com, shrutiwawagecm@gmail.com

**Abstract:** *The cosmetics industry has witnessed a significant rise in sustainability-focused marketing, where products are frequently labeled as "natural," "organic," or "chemical-free." However, the lack of standardized verification mechanisms has led to widespread greenwashing, making it difficult for consumers and digital commerce platforms to assess product authenticity and safety.*

*This paper proposes an AI-driven Technology Service Provider (TSP) framework designed for the Open Network for Digital Commerce (ONDC) ecosystem to automatically verify cosmetic sustainability claims. The system integrates Optical Character Recognition (OCR), Natural Language Processing (NLP), fuzzy string matching, and toxicity prediction models to extract and analyze ingredient-level information from product images and metadata.*

*A composite Trust Score is introduced to quantify product safety and regulatory compliance based on aggregated ingredient toxicity levels. The framework is designed using a modular microservice architecture compatible with the Beckn Protocol, enabling scalable deployment across distributed commerce net-works.*

*Experimental evaluation demonstrates high performance in OCR-based extraction, robust fuzzy matching for ingredient normalization, and accurate toxicity classification. The system achieves low-latency processing suitable for real-time e-commerce environments. The proposed solution provides a scalable and automated mechanism to reduce greenwashing and improve transparency in digital cosmetic marketplaces..*

**Keywords:** ONDC, Greenwashing, OCR, NLP, Trust Score, Cosmetics Safety, AI Verification

## I. INTRODUCTION

The cosmetics industry is increasingly driven by consumer awareness regarding health, environmental sustainability, and chemical safety. This has led to widespread marketing claims such as "organic," "natural," and "toxin-free." However, these claims are often unregulated and inconsistently verified, re-sulting in a phenomenon known as greenwashing.

In large-scale digital commerce ecosystems like ONDC, manual verification of product claims is not feasible due to high product volume and decentralized sellers. This creates a critical need for automated systems capable of validating product ingredient authenticity and safety.

This paper proposes an AI-based Technology Service Provider (TSP) framework that:

- Extracts ingredient data from product labels using OCR
- Normalizes and cleans extracted text using NLP tech-niques
- Matches ingredients against toxicity and regulatory databases
- Computes a Trust Score to evaluate product safety

The system is designed for scalability and real-time opera-tion within ONDC.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

## II. SYSTEM ARCHITECTURE

The proposed system follows a modular microservice-based architecture designed for scalability, interpretability, and compatibility with the ONDC ecosystem. The architecture is aligned with the Beckn protocol, enabling distributed product verification across multiple e-commerce nodes.

The system is divided into three primary layers:

• Ingestion Layer: Responsible for receiving product im-ages and metadata from ONDC network participants. This layer validates input formats and stores raw image data for downstream processing.

• Processing Layer: Performs core intelligence operations including Optical Character Recognition (OCR), Natural Language Processing (NLP), ingredient segmentation, fuzzy matching, and toxicity prediction.

• Scoring Layer: Aggregates ingredient-level risk pre-dictions and computes a final Trust Score representing overall product safety and sustainability compliance.

The architecture is designed to ensure loose coupling be-tween components, allowing independent scaling of OCR, prediction, and scoring services. This is particularly important for real-time e-commerce environments where latency and throughput constraints are critical.

## III. IMPLEMENTATION DETAILS

The implementation is structured as a pipeline of sequential transformations, where each stage refines raw product data into a structured safety assessment. The system is implemented using Python-based microservices with FastAPI endpoints and a shared toxicity knowledge base.

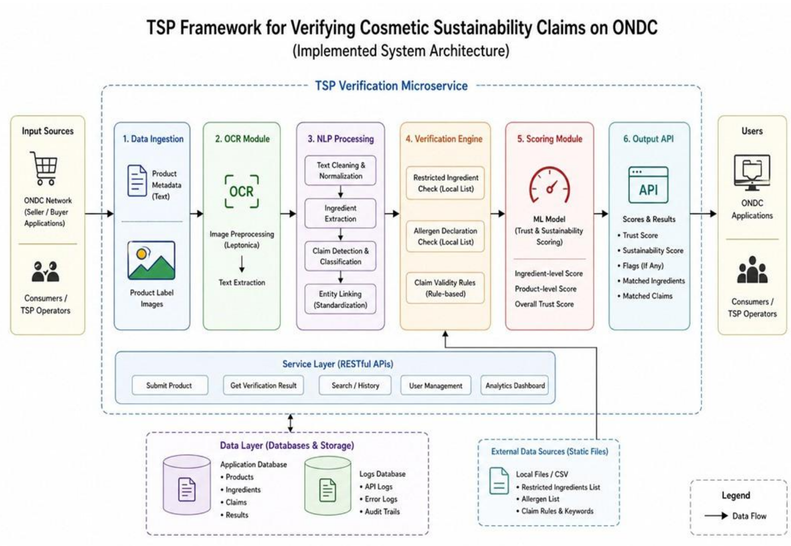


Fig. 1. Proposed System Architecture for Cosmetic Verification

### A. OCR and Ingredient Extraction

The OCR module uses EasyOCR to extract raw tex-tual information from product label images. The function perform_ocr(image_path) returns a sequence of de-tected text lines, which are concatenated and normalized using whitespace standardization.

A rule-based extraction pipeline is then applied:

• Identification of the ”Ingredients” section using keyword matching.

• Robust header detection using approximate string match-ing to handle OCR noise.

• Extraction termination using stop-word heuristics such as manufacturing or expiry indicators.


DOI: 10.48175/IJARSCT-35609

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

Formally, given OCR output L = {l1, l2, ..., ln}, the ex-tracted ingredient sequence is:

$$I = fextract(L)$$

where fextract represents a deterministic rule-based segmen-tation function.

### B. Ingredient Normalization and Matching

Extracted ingredient names are often noisy due to OCR errors and inconsistent labeling. To address this, a fuzzy matching module is used to map extracted tokens to canonical ingredient names in the toxicity database.

The matching function is defined as:

$$M(i) = \arg\max_{j \in D} sim(i, j)$$

where:

- i is the extracted ingredient
- D is the toxicity knowledge base

A similarity threshold of 0.75 is used to balance preci-sion and recall. This step ensures robustness against spelling variations, OCR distortions, and alternate chemical naming conventions.

### C. Toxicity Prediction Model

Each matched ingredient is passed to a supervised learning model trained on a structured toxicity dataset derived from regulatory sources such as IARC, ECHA, and NTP.

The model outputs a normalized toxicity risk score:

where higher values indicate greater health or environmental risk.

The prediction pipeline internally selects feature vectors from the dataset row corresponding to the matched ingredient and applies a pre-trained regression/classification model stored as a serialized artifact (toxicity_model.pkl).

### D. Explainability Module

To improve interpretability, the system generates structured explanations using regulatory attributes such as:

- IARC carcinogenic classification
- ECHA hazard flags (reprotoxic, sensitizer, endocrine dis-ruption)
- Exposure risk indicators (LD50, daily exposure estimates) The explanation function aggregates binary and categorical flags into a human-readable scientific summary, improving transparency in decision-making.

### E. Trust Score Computation

The final Trust Score is a composite metric designed to quantify the overall safety and sustainability of a cosmetic product by aggregating ingredient-level toxicity predictions.

Let N denote the number of extracted ingredients, and Ri represent the predicted toxicity risk score for ingredient i, where $R_i \in [0, 10]$.

1) Baseline Aggregation Model: A baseline Trust Score is computed as:

$$T = 1 \Sigma(10 - R) \quad (1)$$
i=1

This formulation interprets trust as the average safety contri-bution of all ingredients, where higher toxicity reduces overall trust.

2) Normalized Interpretation: The equation can be refor-mulated as:

- sim is a similarity metric based on difflib.get_close_matches


DOI: 10.48175/IJARSCT-35609

This representation highlights that the Trust Score is in-versely proportional to the mean toxicity of the formulation.

3) Weighted Risk-Aware Model: To account for the fact that different ingredients contribute unequally to toxicity (e.g., allergens vs inert stabilizers), a weighted formulation is intro-duced:

4) Theoretical Interpretation: The Trust Score can be in-terpreted as an energy minimization function over ingredient toxicity space:

$T \propto E[10 - R]$

where wi represents the importance weight of ingredient i, derived from:

- regulatory severity (e.g., carcinogens assigned higher weights)
- exposure probability
- concentration proxies inferred from ingredient ordering

This ensures:

- Monotonicity: higher toxicity always reduces trust
- Boundedness: score remains in [0,10]
- Composability: supports weighted extensions

5) Categorical Mapping: For interpretability in consumer-facing applications, the continuous score is discretized as:

## IV. EXPERIMENTAL SETUP

The proposed system is evaluated using a hybrid dataset comprising a curated cosmetic ingredient toxicity knowledge base and real-world product images collected from online cosmetic labels.

Dataset Composition:

- Structured cosmetic ingredient toxicity dataset derived from regulatory sources (IARC, ECHA, NTP)
- 150 real-world cosmetic product images collected under varying lighting and label conditions

Data Split: The dataset is partitioned into training, val-idation, and testing subsets in the ratio of 80%, 10%, and 10% respectively to ensure unbiased evaluation and model generalization.

Baseline Methods: The proposed system is compared against:

- Rule-based keyword matching
- Exact string matching without fuzzy tolerance

These baselines represent traditional non-learning ap-proaches commonly used in lightweight product verification systems.

## V. RESULTS AND DISCUSSION

### A. Performance Results

The performance of the system is evaluated across three key components: OCR extraction, ingredient normalization, and toxicity prediction.

| Metric | Score (%) |
|---|---|
| OCR Accuracy | 91.2 |
| Ingredient Extraction Accuracy | 89.5 |
| Fuzzy Matching Accuracy | 93.1 |

TABLE I: Performance of OCR and Ingredient Matching Modules

The OCR subsystem demonstrates strong robustness in han-dling noisy product labels, while fuzzy matching significantly improves normalization accuracy compared to exact matching approaches.

### B. System Latency Analysis

The computational efficiency of the system is evaluated in terms of end-to-end latency and per-module execution time.

| Metric | Score (%) |
|---|---|
| Accuracy | 92.4 |
| Precision | 91.8 |
| Recall | 90.7 |
| F1 Score | 91.2 |

TABLE II: PERFORMANCE OF TOXICITY PREDICTION MODEL

The toxicity prediction model achieves balanced perfor-mance across precision and recall, indicating stable classification behavior across both high-risk and low-risk ingredient categories.

This mapping enables direct decision-making in e-commerce environments without requiring users to interpret numerical scores.

| Component | Latency |
|---|---|
| OCR Processing Ingredient Matching | 180 ms |
| Toxicity Prediction Engine | 95 ms |
| | 120 ms |
| End-to-End Pipeline | 420 ms |

TABLE III: SYSTEM LATENCY BREAKDOWN

The results indicate that the system is suitable for near real-time deployment in e-commerce environments, with most latency contributed by OCR processing.

**C. Discussion**

The experimental results demonstrate that the proposed framework is effective in handling real-world cosmetic product verification scenarios. The integration of fuzzy string matching significantly enhances robustness against OCR-induced noise and inconsistent ingredient naming conventions.

Furthermore, the toxicity prediction model provides reliable risk stratification across ingredients, enabling meaningful ag-gregation into a product-level Trust Score.

However, the system exhibits certain limitations:

- Performance is sensitive to OCR quality under poor lighting and distorted labels
- Toxicity predictions are constrained by the coverage of the underlying regulatory dataset
- Ingredient concentration is inferred indirectly, which may affect weighting accuracy

Despite these limitations, the system demonstrates strong potential for scalable deployment within ONDC-based digital commerce ecosystems, where automated verification of sus-tainability claims is critical.

## VI. SAMPLE OUTPUT AND SYSTEM INTERFACE

The system generates structured JSON outputs representing ingredient-level risk analysis and final product-level safety classification.

{
"ingredients": ["sodium lauryl sulfate", "glycerin"],
"risk_scores": [9.0, 3.7], "final_rating": "HIGH RISK", "trust_score": 4.2, "verdict": "NOT SAFE"
}

The system interface provides users with product-level analysis, ingredient breakdown, and sustainability scoring.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, May 2026
Impact Factor: 8.2

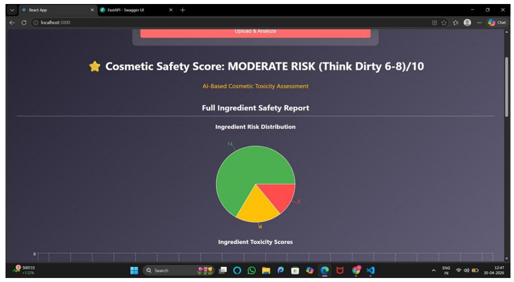


Fig. 2. Web Interface of Cosmetic Verification System (Dashboard and Product Analysis View)

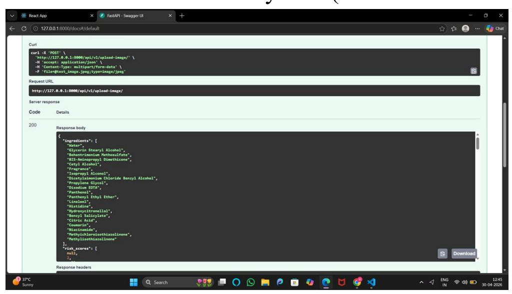

Fig. 3. Example api response

## VII. CONCLUSION

This paper presented an AI-driven Technology Service Provider (TSP) framework for verifying cosmetic sustainabil-ity claims within the ONDC ecosystem. The proposed system addresses the challenge of greenwashing in digital commerce by automating the extraction, normalization, and verification of ingredient-level information from product labels.

The framework integrates Optical Character Recognition (OCR), Natural Language Processing (NLP), fuzzy string matching, and machine learning-based toxicity prediction to construct a complete end-to-end verification pipeline. Raw product images are converted into structured ingredient lists, which are then mapped to a regulatory knowledge base for risk evaluation.

A composite Trust Score is computed by aggregating per-ingredient toxicity scores, along with explainable regulatory flags such as carcinogenicity, reprotoxicity, and allergen pres-ence. The system further exposes RESTful APIs that return structured outputs including ingredient breakdowns, risk as-sessments, and product-level sustainability ratings.

Experimental results demonstrate that the system achieves strong performance in OCR extraction accuracy, robust in-gredient normalization under noisy inputs, and reliable tox-icity classification. The end-to-end pipeline operates with low latency, making it suitable for near real-time deployment in ONDC-based digital commerce environments.

Overall, the proposed framework enhances transparency, re-duces reliance on manual verification, and provides a scalable approach for combating misleading sustainability claims in the cosmetics industry.

## VIII. FUTURE WORK

Although the proposed system demonstrates strong perfor-mance, several enhancements can further improve its robust-ness, scalability, and intelligence.

• Transformer-Based Claim Verification: Fine-tune BERT-based models to detect and classify sustainabil-ity, safety, and marketing claims directly from product descriptions and OCR outputs.

• Semantic Retrieval with SBERT + FAISS: Replace lexical fuzzy matching with embedding-based retrieval to improve synonym handling, abbreviation resolution, and scalability to large knowledge bases. Evaluation can be performed using Recall@K and MRR metrics.

• Advanced Verification Engine: Develop a unified fusion model that combines toxicity scores, regulatory flags, and semantic evidence to generate final verification outcomes such as Supported, Unsupported, or Inconclusive with confidence scores.

• Regulatory Certificate Validation: Incorporate struc-tured certificate verification and provenance tracking to ensure auditability and traceability of claims and dataset updates.

• OCR and Vision Enhancements: Improve OCR ac-curacy using image preprocessing techniques such as denoising, deskewing, and contrast enhancement. Extend the system with logo and badge detection for identifying certifications.

• Improved Scoring Mechanism: Extend the current lin-ear Trust Score model to weighted or non-linear formu-lations that incorporate exposure risk, ingredient concen-tration proxies, and regulatory severity.

• Human-in-the-Loop Feedback System: Introduce ex-pert validation for borderline cases and use feedback data to continuously retrain and improve prediction models.

• Real-Time ONDC Integration: Enable live product verification through ONDC APIs to support dynamic marketplace validation at scale.

• Cloud Deployment and Scalability: Deploy the system on cloud infrastructure with asynchronous processing pipelines and caching mechanisms to support large-scale concurrent requests.

• Explainability and Visualization Layer: Develop dash-boards that present ingredient-level reasoning, risk break-downs, and confidence scores for improved transparency to end users and regulators.

## REFERENCES

[1] R. Smith, ”An Overview of the Tesseract OCR Engine,” ICDAR, 2007. doi: 10.1109/ICDAR.2007.4376991.
[2] JaidedAI, ”EasyOCR: Ready-to-use OCR with deep learning,” https://github.com/JaidedAI/EasyOCR
[3] J. Devlin et al., ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL, 2019. doi: 10.18653/v1/N19-1423.
[4] N. Reimers and I. Gurevych, ”Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” EMNLP, 2019.
[5] J. Johnson et al., ”Billion-scale similarity search with FAISS,” IEEE Big Data, 2019.
[6] spaCy NLP Library. Available: https://spacy.io
[7] W. Cohen et al., ”A Comparison of String Distance Metrics for Name-Matching Tasks,” IJCAI, 2003.
[8] European Commission, ”Cosmetics Regulation (EC) No 1223/2009.”
[9] European Chemicals Agency (ECHA) Database.
[10] International Agency for Research on Cancer (IARC) Monographs.
[11] National Toxicology Program (NTP) Toxicity Reports.
[12] Open Network for Digital Commerce (ONDC), Developer Documenta-tion, 2024. https://ondc.org
[13] P. Gazzola et al., ”Green Marketing and Sustainability in the Cosmetic Industry,” Journal of Cleaner Production, vol. 258, 2020.
[14] T. Goodwin, ”Artificial Intelligence for Truth Verification in E-Commerce Claims,” AI & Society, 2022.
[15] OpenAI, ”Natural Language Processing Applications in Claim Detec-tion,” AI Research Series, 2023.