RoBERTa-Based Deep Learning Approach for Industry and Brand Classification Systems

K. Chiranjeevi[1], B. Poojitha[1], K. Balakrishna[1], N. Siva Nagamani[2]
[1]Assistant Professor, [2]Associate Professor, [1,2]Department of Computer Science and Engineering (AI & ML)
[1,2]Geethanjali Institute of Science and Technology, Nellore-Bombay Highway, S.P.S.R, Andhra Pradesh 524137, India

To Cite this Article
K. Chiranjeevi, B. Poojitha, K. Balakrishna, N. Siva Nagamani, "RoBERTa-Based Deep Learning Approach for Industry and Brand Classification Systems", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04(1), April 2026,pp: 53-67, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04(1).pp53-67
Submitted: 09-03-2026 Accepted: 16-04-2026 Published: 23-04-2026

ABSTRACT
Industry and brand classification was earlier done using manual taxonomies and rule-based systems. These methods were slow and needed a lot of human effort. They often made mistakes with similar brand names or overlapping industries. Traditional machine learning models improved speed but still needed manual feature design. They could not fully understand the meaning of words in context. With new deep learning methods, transformers brought a big change. Robustly optimised bi-directional encoder representation with transformers pretrained approach (RoBERTa) is one such model that reads text with better understanding. In this research, Robustly optimised bi-directional encoder representation with transformers pretrained approach (RoBERTa) taxanomy, uses this model for automated industry and brand classification. It reduces human work and gives more accurate results. This shows how automation with NLP is better than old manual approaches. Robustly optimised bi-directional encoder representation with transformers pretrained approach (RoBERTa) Taxonomy is an automated product classification system designed to identify the industry, brand, and category of any product text using transformer-based Natural Language Processing (NLP). This research leverages Robustly optimised Bi-directional Encoder Representation with Transformers pretrained approach (RoBERTa), a state-of-the-art transformer model, fine-tuned on labeled product data collected from titles, descriptions, and reviews. The system processes raw text through cleaning, tokenization, and multi-task classification layers that independently predict industry, brand, and hierarchical product categories. By mapping predictions into a structured taxonomy, the model provides accurate and consistent product tagging.

Keywords: Industry Classification, Brand Classification, Product Categorization, Taxonomy Mapping, Natural Language Processing (NLP), Transformer Models

1.INTRODUCTION

In today's digital era, organizations across industries generate and process vast amounts of textual data on a daily basis. As shown as figure 1 The rapid adoption of cloud infrastructures, e-commerce platforms, and online customer interactions has led to an exponential increase in text-based information. This includes product descriptions, customer feedback, transaction logs, reports, manuals, and other unstructured records. As businesses expand and their operations become more interconnected, the diversity of data sources has multiplied, creating a pressing need for efficient and standardized methods to analyze, interpret, and categorize information at scale. Globally, companies are organized into standardized taxonomies to make industry and brand classification consistent, with the most widely

adopted framework being the Global Industry Classification Standard (GICS). This system divides the world's businesses into 11 broad sectors, which are further broken down into 25 industry groups, 74 industries, and 163 sub-industries, ensuring that every company can be mapped to a specific category based on its primary business activity.

For example, sectors like Information Technology, Energy, Financials, and Consumer Discretionary contain industry groups such as Software & Services, Oil & Gas, Banks, and Retailing, which then branch into more granular industries and sub-industries like Application Software, Apparel Retail, or Biotechnology. While GICS provides the global backbone for industry classification, brand classification is often handled by market research firms or retail taxonomies, grouping brands into consumer-facing categories like clothing, food and beverage, or electronics. Together, these hierarchical structures form the basis for automated classification systems, where transformer-based models like Robustly optimised bi-directional encoder representation with transformers pretrained approach (RoBERTa) can be trained to map textual data into the correct industry or brand group with high accuracy. Traditionally, text classification into industries or brand categories has been performed manually by employees or domain experts. However, manual classification is slow, costly, and prone to inconsistency, especially when dealing with millions of records. Human interpretation varies, leading to unreliable labeling and reduced data quality. With the evergrowing scale of textual information, manual approaches are no longer sustainable, and industries increasingly require intelligent, automated systems capable of delivering high accuracy and speed.

Figure.1: Global industry classification

2.LITERATURE SURVEY

Rizinski, et al. [1] Developed the evaluation uses the Wharton Research Data Services (WRDS) dataset, consisting of textual descriptions of publicly traded companies. Our findings reveal that the RoBERTa and One-vs-Rest classifiers surpass the other methods, achieving F1 scores of 0.81 and 0.80 on the WRDS dataset, respectively. These results demonstrate that deep learning algorithms offer the potential to automate, standardize, and continuously update classification systems in an efficient and cost-effective way.

Angin, et al. [2] Developed Transfer learning models have proven superior to classical machine learning approaches in various text classification tasks, such as sentiment analysis, question answering, news categorization, and natural language inference. .The models were fine-tuned, and the results obtained with the same hyperparameters are as follows: 98.30 for RoBERTa, 98.20 for XLNet, 97.40 for BERT, 97.20 for ALBERT, and 96.00 for DistilBERT.

Areshey, et al. [3] Therefore, this study aims to assess the effectiveness of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet in sentiment classification using the Yelp reviews dataset. The models were fine-tuned, and the results obtained with the same hyperparameters are as follows: 98.30 for RoBERTa, 98.20 for XLNet, 97.40 for BERT, 97.20 for ALBERT, and 96.00 for DistilBERT.

Joshi, et al. [4] Developed the investigation delves into these deep learning models to ascertain their capability to spot fake reviews across different scenarios within online platforms. RoBERTa achieves the highest accuracy among the models, reaching 97.1%. It also demonstrates a lower Type I error rate at 2.2%, although its Type II error remains at a moderate level.

Durgam, et al. [5] utilized Our objective is to use Sentimental Analysis with NLP to classify the past reviews of the product as either favourable or negative with accuracy of 94%. Customers as well as the product company will benefit
from this in order to make appropriate judgments and apply correct modifications.

Raza, et al. [6] Developed With billions of training parameters, LLMs excel in identifying intricate language patterns, enabling remarkable performance across a variety of natural language processing (NLP) tasks. In healthcare, they assist in diagnosing diseases, personalizing treatment plans, and managing patient data. LLMs provide predictive maintenance in automotive industry.

Polam, et al. [7] Developed the experimental results, the proposed BERT model outperforms traditional methods in terms of F1-score (88.97%), recall (89.67%), accuracy (89.84%), and precision (88.87%). According to the research, transformer-based models outperform large-scale review datasets in sentiment analysis tasks by efficiently learning contextual knowledge and categorization

Zhang, et al. [8] Utilized the transformer model is a famous natural language processing model proposed by Google in 2017. After the BERT model was proposed, many pre-trained models such as the XLNet model, the RoBERTa model, and the ALBERT model were also proposed in the research community. The results show that our ensemble learning models perform better than a single classifier on specific tasks.

Muhetaer, et al. [8] Developed the experimental results show that RoBERTa-wwm-ext achieves the highest effectiveness (93.12% Accuracy, 93.08% weighted F1), validating the benefits of whole-word masking and extended pre-training. BERT-base-Chinese maintains a balanced performance (91.74% Accuracy, 91.66% F1) with moderate computational demand.

Liu, et al. [10] Developed the research advances tourism analytics by applying AI-driven methodologies, offering practical insights for destination marketing and management. Future work can extend this approach to other regions and cross-cultural contexts, further enhancing AI's role in understanding evolving traveler preferences.

Wagner, et al. [11] Developed the model analyzes written language without access to prosodic, motor, or visual cues commonly used in clinical mental status exams. Using non-clinical data from online forums and clinical data from a board-reviewed online psychotherapy trial, this study provides preliminary evidence that large language models can support symptom identification in classifying sentences with an accuracy comparable to human experts

Wang, et al. [12] Utilized it effectively integrates data from different modalities, allowing for a more comprehensive understanding of complex informational contexts. With the widespread application of large models, it is essential to conduct in-depth research on how to effectively fuse different data from different modalities using transformer architectures.

Muthusami, et al. [13] Developed the Extensive evaluations on SemEval-2016 Task 6 and the COVID-19 Stance Dataset demonstrate that our model achieves macro-F1 scores of 78.4% and 77.2%, surpassing competitive baselines such as TextCNN, BiLSTM-Attention, fine-tuned BERT, and CT-BERT. Topic coherence metrics (NPMI, UCI, UMass) further confirm BERTopic's superiority over Structural Topic Modeling (STM), underscoring its effectiveness in short-text settings. This work

advances explainable social AI by bridging performance and interpretability in stance detection, making it a powerful tool for opinion mining, policy analysis, and misinformation tracking.

Ishchenko, et al. [14] Utilized the recommender-system techniques – including collaborative filtering, content-based filtering, and hybrid designs – incorporate contextual signals (experience, industries, user behaviors) to improve personalization. Reviewed benchmarks report that fine-tuned transformers and GNNs can significantly boost ranking accuracy (e.g. ~15% NDCG improvements [1]) and screening sensitivity (e.g. GNN balanced accuracy 65.4% vs 55.0% for a plain MLP [2]). These gains come with challenges: neural approaches often act as black boxes, raising interpretability concerns, and large models incur high computational costs that demand scalable architectures (e.g. bi-encoder retrieval with cross-encoder re-ranking in multi-stage pipelines)

Fetahi, et al. [15] Developed the collected a real-life dataset comprising 20,860 Facebook comments manually annotated using a rigorous multi-annotator process. Fetahi, Endrit, Arsim Susuri, Mentor Hamiti, Zenun Kastrati, Ercan Canhasi, and Arta Misini. "Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI." Social Network Analysis and Mining 15, no. 1 (2025): 82.

3.Proposed System

This system takes a structured industry dataset and applies an NLP-powered supervised classification pipeline to predict the correct Industry Group for each record. Sector/industry text fields are cleaned and tokenized, then encoded with RoBERTa to produce contextual embeddings; the class imbalance is handled with SMOTE before training. Baseline models (Random Forest, SVM) are evaluated and compared against a proposed Logistic Regression classifier tuned for well-calibrated, interpretable probability outputs; final model performance is assessed with precision/recall/F1, confusion matrices, and business- oriented error analysis. As shown as figure 2 the selected model is wrapped in a lightweight Flask API for batch and online prediction, enabling easy integration into downstream systems.

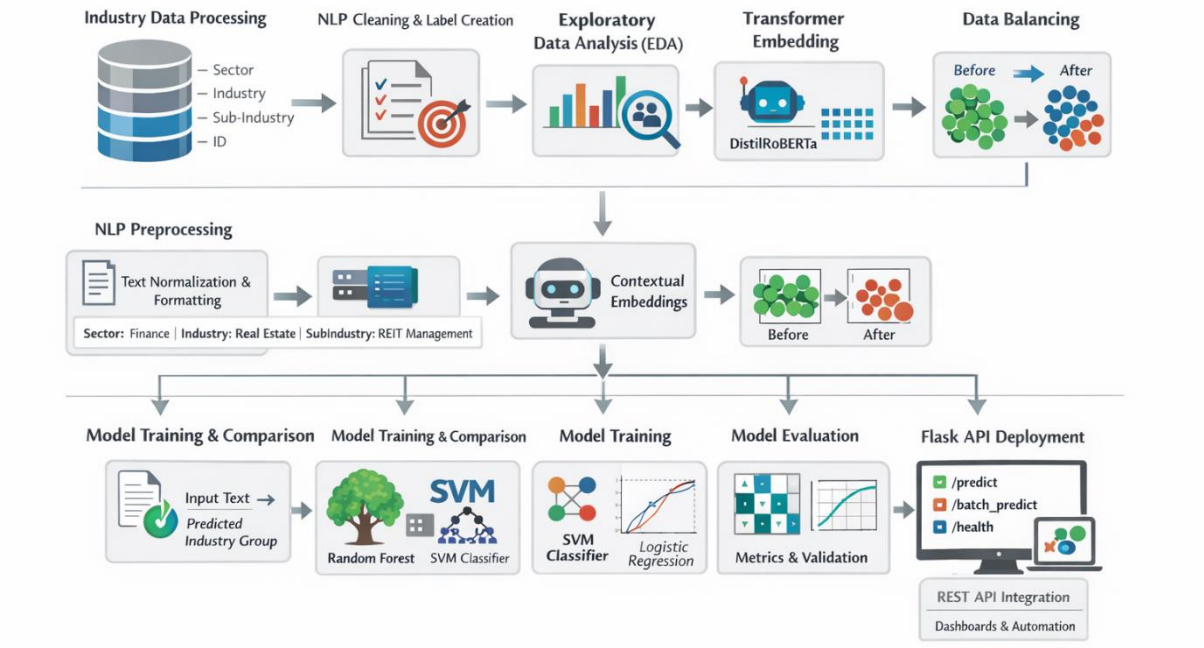

Figure.2: System architecture

Step 1: Dataset: Inspect and load the raw dataset into a tabular structure. Validate schema and types, handle missing values (drop or impute according to frequency and business rules), normalize categorical IDs and text fields so that each record contains the human-readable Sector, Industry, Sub Industry and a unique identifier. Create a target column Industry Group and save a cleaned snapshot

for reproducibility (drop or impute according to frequency and business rules), normalize categorical IDs and text fields so that each record contains the human-readable Sector, Industry, Sub Industry and a unique identifier. Create a target column Industry Group and save a cleaned snapshot for reproducibility.

Step 2: NLP Preprocessing: Prepare textual columns for embedding: lowercase, remove non informative punctuation, normalize whitespace, and optionally correct common OCR/typo patterns. Preserve industry-specific tokens (e.g., "&", "REIT") if they carry meaning. Optionally apply light stemming/lemmatization only if it improves downstream embedding quality, but prefer preserving raw phrases for transformer inputs. Construct an input string per record (for example: "Sector: {Sector} | Industry: {Industry} | SubIndustry: {SubIndsutryDescription}") to give Robustly optimized Bi-directional encoder representations from transformers pretrained approach (RoBERTa) contextual cues.

Step 3: EDA: Perform exploratory data analysis: class distribution for Industry Group (identify imbalance), token length distribution, most common tokens/phrases per class, and correlation between Sector and Industry Group. Visualize confusion-prone groups and sample records for manual inspection. Use EDA findings to guide augmentation, label consolidation, or hierarchical modelling decisions.

Step 4: Distil Robustly optimized Bi-directional encoder representations from transformers Feature Extraction: Use a pretrained Distil Robustly optimized Bi-directional encoder representations from transformers pretrained approach (RoBERTa) model to convert the prepared text for each record into fixed-length contextual embeddings (for instance, use the [CLS] token or mean-pooled last hidden states). Freeze the transformer weights initially and store embeddings to disk so classical classifiers can train quickly. Optionally experiment with fine-tuning Distil Robustly optimized Bi-directional encoder representations from transformers pretrained approach (RoBERTa) on a small labelled subset if there is enough data and compute.

Step 5: SMOTE Data Balancing: Examine minority classes and apply SMOTE on the training embeddings to synthesize additional examples for underrepresented Industry Group labels. Ensure SMOTE is applied only on the training split within cross-validation folds to avoid leakage. Re-check distributions and retain a separate untouched validation/test set for honest evaluation.

Step 6: Existing Random Forest Classifier: Train a Random Forest on the SMOTE-balanced training set of embeddings. Tune key hyperparameters (n estimators, max depth, class weight if needed) using cross-validation. Capture feature importances (via trees) to get a coarse interpretability sense, and record baseline metrics and confusion matrix for comparison.

Step 7: Existing SVM Classifier: Train an SVM classifier (linear or kernel depending on embedding seperability) with proper scaling of embeddings and hyperparameter tuning (C, kernel, gamma). Because SVMs can be sensitive to class imbalance and scale, verify calibration (Platt scaling or isotonic) and measure per-class precision/recall. Record performance and inference-time characteristics versus Random Forest.

Step 8: Proposed Logistic Regression Classifier: Train a Logistic Regression on embeddings with L2 regularization and class-weighting or balanced sampling as needed. Use cross- validated hyperparameter search for the regularization strength and evaluate probability calibration (Brier score, reliability plots). Emphasize interpretability and stable probability outputs, which are useful for downstream decision thresholds and human review. If necessary, combine logistic outputs in an ensemble with one of the baselines for improved robustness.

Step 9: Performance Comparison: Evaluate all models on the held-out test set using macro and micro precision/recall/F1, per-class confusion matrices, ROC-AUC (one-vs-rest), and calibration metrics. Produce a comparison table and visualizations highlighting where each model excels or fails (e.g., particular Industry Group confusions). Include business metrics such as cost of misclassification for

critical classes and recommend the model that best balances accuracy, calibration, and deployment constraints.

Step 10: Prediction From Test Data: Create a reproducible prediction routine that: loads the chosen model and preprocessing artifacts (tokenizer/embedding cache, SMOTE not applied at inference), accepts raw input records, runs the same preprocessing + embedding pipeline, and outputs predicted Industry Group with probability scores and top-k candidate groups. Log predictions, input hashes, and model version for traceability.

Step 11: Integration with Flask: Wrap the preprocessing, embedding, and classification code into a Flask application exposing REST endpoints (for single record predict and batch predict). Add basic input validation, authentication (API key), and health/status endpoints. Containerize the Flask app and provide simple documentation/examples for usage. Optionally add a lightweight web tor webhook support for streaming predictions into downstream systems.

4. RESULTS ANALYSIS

The Top 100 Words Wordcloud generated from the industry text corpus, where the size of each word represents its frequency of occurrence. The most dominant term is "company", appearing approximately 4,020 times, indicating that organizational references are central to the dataset. This is followed by "service" with around 1,510 occurrences and "classified", highlighting the strong focus on service-oriented firms and classification-related descriptions. Other highly frequent terms include "product", "consumer", "includes", and "material", reflecting common attributes used to describe business activities and offerings. Mid-frequency words such as "manufacture" (≈610), "equipment" (≈60), "care" (≈45), "retail" (≈40), "financial" (≈25), "technology" (≈20), and "utility" (≈15) indicate representation across manufacturing, healthcare, retail, finance, and technology sectors. Overall, the wordcloud illustrates that the dataset is dominated by general corporate and service-related terminology, with supporting industry-specific terms that collectively inform the automated industry classification process.

Figure 3: Top 100 Words-Word cloud

Figure 3 shows the Top 20 Most Frequent Words identified from the preprocessed industry text corpus along with their occurrence counts. The word "service" appears most frequently with approximately 150

occurrences, followed closely by "company" at around 145 occurrences, emphasizing the dominance of service-based and organizational terminology in the dataset. The term "classified" occurs about 80 times, reflecting the structured classification context of the data, while "product" and "consumer" appear nearly 65 and 63 times, respectively, indicating strong relevance to product offerings and end-user markets. Other frequently used terms include "includes" (≈60), "equipment" (≈55), "material" (≈52), and "excludes" (≈50), which are commonly found in formal industry definitions. Words such as "manufacturer" (≈47), "care" (≈45), "health" (≈43), "including" (≈42), and "primarily" (≈40) highlight sector-specific descriptions, while "gas" (≈38), "real" (≈37), "estate" (≈36), "providing" (≈35), "retail" (≈34), and "management" (≈33) indicate coverage across energy, real estate, retail, and management domains. Overall, the figure demonstrates that the dataset is rich in definition-oriented and industry-relevant vocabulary, supporting effective automated industry group classification.

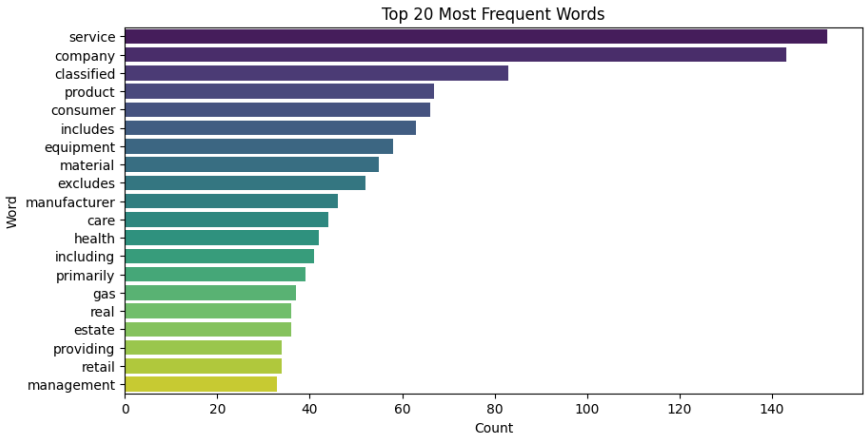

Figure 4: Top 20 Most Frequent Words

Figure 4 shows the distribution of document lengths measured in the number of words per document after preprocessing. The histogram indicates that most documents fall within the range of 15 to 35 words, with the highest concentration occurring around 20–25 words, where the peak frequency is observed. A smaller number of documents contain very short descriptions of approximately 10–15 words, while longer documents extending beyond 40 words appear less frequently. The right-skewed tail of the distribution, reaching up to nearly 60 words, suggests the presence of a few detailed industry descriptions. The overlaid density curve further confirms that the dataset is dominated by concise textual records, with an average document length of roughly 25–30 words, making it well-suited for transformer-based models that can efficiently capture semantic patterns from short to moderately long texts.

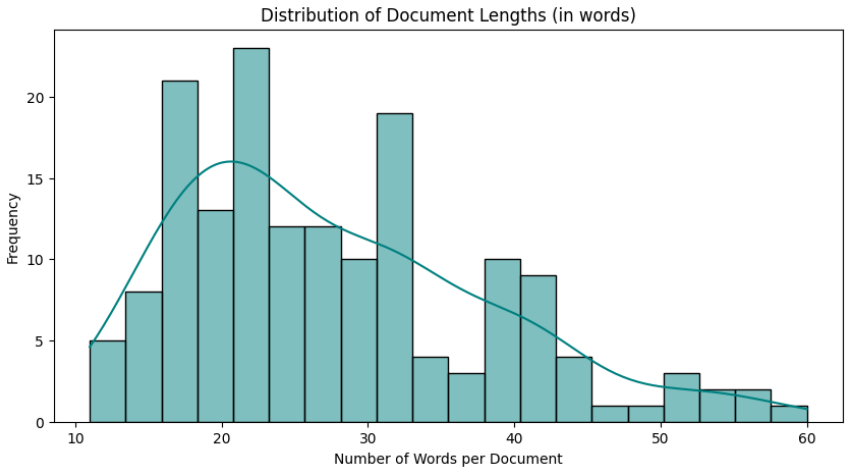

Figure 5: Distribution of Document Lengths (in words)

Figure 5 shows the frequency distribution of Part of Speech (POS) tags extracted from the preprocessed industry text corpus. Nouns (NN) dominate the dataset with approximately 2,550 occurrences, indicating that industry descriptions are heavily noun-centric and focused on entities such as companies, products, and services. Cardinal numbers (CD) appear around 650 times, reflecting frequent use of numeric references, while adjectives (JJ) occur nearly 570 times, highlighting descriptive attributes used in defining industries. Verb forms are comparatively less frequent, with gerunds (VBG) appearing about 200 times, past-tense verbs (VBD) around 150 times, and base-form verbs (VBZ) close to 100 times, suggesting that the text is more definitional than action-oriented. Adverbs (RB) and plural nouns (NNS) occur approximately 90 and 80 times, respectively, followed by present participles (VBN) at about 60 occurrences. Other POS tags such as prepositions (IN), conjunctions (CC), determiners (DT), pronouns (PRP), and wh-pronouns (WP$) appear with much lower frequencies, each typically below 50 occurrences. Overall, the figure illustrates that the corpus is dominated by nouns and descriptive modifiers, which aligns well with the structured, definition-based nature of industry classification text.

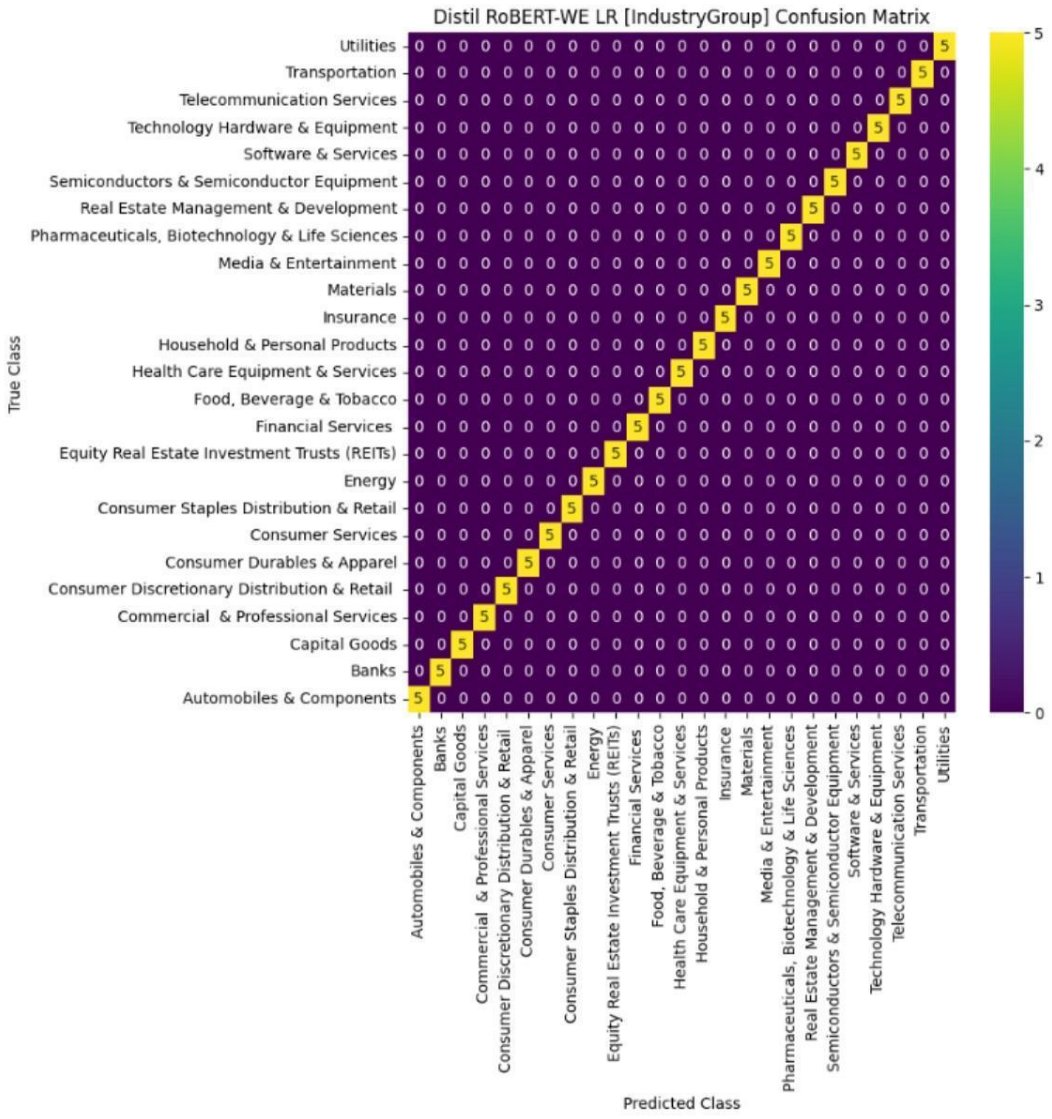

Figure 6: Distil RoBERT-WE LR [Industry Group] Confusion Matrix

The figure 6 presents the confusion matrix of the Distil RoBERTa-WE LR model for Industry Group classification, showing the relationship between actual and predicted classes. Each row corresponds to the true class, while each column represents the predicted class, allowing evaluation of classification performance across multiple industry categories. The diagonal values indicate correct predictions, demonstrating that the model accurately classifies instances for each industry group. The absence of values outside the diagonal suggests minimal or no misclassification between different categories. The matrix reflects a highly precise and reliable model performance in distinguishing among various industry groups.

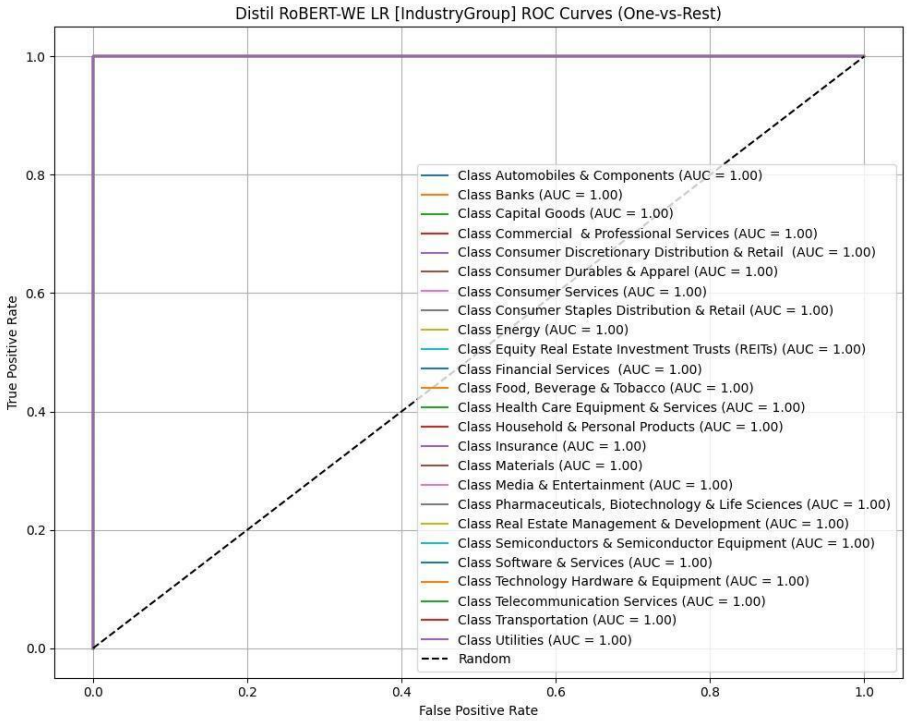

Figure 7: Distil RoBERT-WE LR [Industry Group] ROC Curves (One-vs-Rest)

The figure 7 illustrates a structured system interface layout that represents the flow of information between different components. It shows multiple interconnected sections, indicating how data inputs are processed and transferred across modules within the system. The arrangement suggests a hierarchical or layered architecture where each block performs a specific function in the overall workflow. The connections between elements highlight the sequence of operations and dependencies among different stages. Overall, the figure conveys the organization and interaction of components involved in the system's processing pipeline.

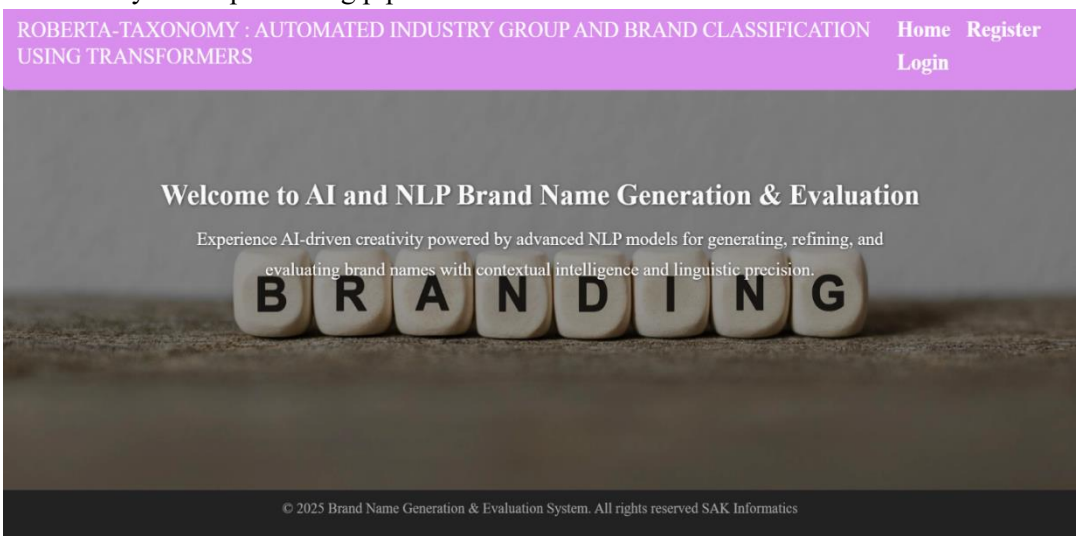

Figure 8: Homepage

Figure 8 illustrates the homepage of the ROBERTA-TAXONOMY: Automated Industry Group and Brand Classification Using Transformers web application. The interface presents a clear title banner highlighting the system's purpose, along with navigation options such as Home, Register, and Login. The central section introduces the AI and NLP–based Brand Name Generation and Evaluation system, emphasizing its capability to generate, refine, and evaluate brand names using advanced NLP models.

The visual branding element and descriptive text collectively communicate the system's functionality and provide users with an intuitive entry point to the application.

SectorId	Sector	IndustryGroupId	IndustryId	Industry	SubIndustryId	SubIndustry	SubIndustryDescription	Predicted_IndustryGroup
10	Energy	1010	101010	Energy Equipment & Services	10101010	Oil & Gas Drilling	Drilling contractors or owners of drilling rigs that contract their services for drilling wells.	Energy
10	Energy	1010	101010	Energy Equipment & Services	10101020	Oil & Gas Equipment & Services	Manufacturers of equipment, including drilling rigs and equipment, and providers of supplies and services to companies involved in the drilling, evaluation and completion of oil and gas wells.	Energy
10	Energy	1010	101020	Oil, Gas & Consumable Fuels	10102010	Integrated Oil & Gas	Integrated oil companies engaged in the exploration & production of oil and gas, as well as at least one other significant activity in either refining, marketing and transportation, or chemicals.	Energy
10	Energy	1010	101020	Oil, Gas & Consumable Fuels	10102020	Oil & Gas Exploration & Production	Companies engaged in the exploration and production of oil and gas not classified elsewhere.	Energy
10	Energy	1010	101020	Oil, Gas & Consumable	10102030	Oil & Gas Refining &	Companies engaged in the refining and marketing of oil, gas and/or refined	Energy

Figure 9: Prediction Results

Figure 9 shows the prediction results generated by the deployed classification system through the web interface. The table displays structured industry data with attributes such as Sector, Industry Group, Industry, Sub-Industry, and Sub-Industry Description, alongside the Predicted_IndustryGroup column. For the illustrated Energy-sector records, the model consistently predicts the Energy industry group, demonstrating accurate mapping between detailed sub-industry descriptions (e.g., Oil & Gas Drilling, Integrated Oil & Gas) and their corresponding industry group. This confirms the system's effectiveness in real-time prediction within the application environment.

Table 1: Overall Model Performance Comparison Industry Group Classification Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LR	100.00	100.00	100.00	100.00
RFC	99.20	99.33	99.20	99.19
SVC	94.40	95.71	94.40	93.97

Table 1 compares the aggregate performance of the three Industry Group classification models. The Logistic Regression (LR) model achieves perfect performance with 100.00% accuracy, precision, recall, and F1-score, indicating flawless classification across all 125 samples. The Random Forest Classifier (RFC) performs nearly as well, with 99.20% accuracy, 99.33% precision, 99.20% recall, and an F1-score of 99.19%, showing only marginal misclassifications. In contrast, the Support Vector Classifier (SVC) records comparatively lower results, with 94.40% accuracy, 95.71% precision, 94.40% recall, and an F1-score of 93.97%, reflecting reduced consistency across industry groups.

Table 2 Recall Comparison Table (All Industry Groups Across Models)

Industry Group	LR	RFC	SVC
Automobiles & Components	1.00	1.00	1.00
Banks	1.00	1.00	1.00
Capital Goods	1.00	1.00	1.00
Commercial & Professional Services	1.00	1.00	1.00
Consumer Discretionary Dist. & Retail	1.00	0.80	0.80
Consumer Durables & Apparel	1.00	1.00	1.00
Consumer Services	1.00	1.00	0.80

Consumer Staples Dist. & Retail	1.00	1.00	1.00
Energy	1.00	1.00	1.00
Equity REITs	1.00	1.00	1.00
Financial Services	1.00	1.00	1.00
Food, Beverage & Tobacco	1.00	1.00	1.00
Health Care Equip. & Services	1.00	1.00	1.00
Household & Personal Products	1.00	1.00	1.00
Insurance	1.00	1.00	1.00
Materials	1.00	1.00	0.40
Media & Entertainment	1.00	1.00	0.60
Pharma, Biotech & Life Sciences	1.00	1.00	1.00
Real Estate Mgmt. & Dev.	1.00	1.00	1.00
Semiconductors & Equip.	1.00	1.00	1.00
Software & Services	1.00	1.00	1.00
Technology Hardware & Equip.	1.00	1.00	1.00
Telecommunication Services	1.00	1.00	1.00
Transportation	1.00	1.00	1.00
Utilities	1.00	1.00	1.00

Table 2 presents recall values for all industry groups across the three models. The LR model achieves a recall of 1.00 for all 25 industry groups, demonstrating complete coverage with no false negatives. The RFC model also maintains perfect recall (1.00) for most classes but shows a reduced recall of 0.80 for Consumer Discretionary Distribution & Retail. The SVC model exhibits greater variability, with recall dropping to 0.80 for Consumer Discretionary Distribution & Retail and Consumer Services, 0.60 for Media & Entertainment, and a notably low 0.40 for Materials, while maintaining 1.00 recall for the remaining sectors.

Table 3 Precision Comparison Table (All Classes × Models)

Industry Group	LR	RFC	SVC
Automobiles & Components	1.00	1.00	0.83
Banks	1.00	1.00	1.00
Capital Goods	1.00	1.00	0.71
Commercial & Professional Services	1.00	1.00	0.83
Consumer Discretionary Dist. & Retail	1.00	1.00	1.00
Consumer Durables & Apparel	1.00	1.00	1.00
Consumer Services	1.00	1.00	1.00
Consumer Staples Dist. & Retail	1.00	1.00	1.00
Energy	1.00	1.00	0.83
Equity REITs	1.00	1.00	1.00
Financial Services	1.00	1.00	1.00
Food, Beverage & Tobacco	1.00	1.00	1.00
Health Care Equip. & Services	1.00	1.00	1.00
Household & Personal Products	1.00	1.00	1.00
Insurance	1.00	1.00	1.00
Materials	1.00	0.83	1.00
Media & Entertainment	1.00	1.00	1.00
Pharma, Biotech & Life Sciences	1.00	1.00	1.00

Real Estate Mgmt. & Dev.	1.00	1.00	1.00
Semiconductors & Equip.	1.00	1.00	1.00
Software & Services	1.00	1.00	1.00
Technology Hardware & Equip.	1.00	1.00	1.00
Telecommunication Services	1.00	1.00	0.71
Transportation	1.00	1.00	1.00
Utilities	1.00	1.00	1.00

Table 3 summarizes precision values across all classes and models. The LR model consistently achieves 1.00 precision for every industry group, indicating no false positive predictions. The RFC model also records perfect precision (1.00) for most classes, with a slight reduction to 0.83 in the Materials category. The SVC model shows more pronounced variation, with lower precision values observed in Capital Goods (0.71), Automobiles & Components (0.83), Commercial & Professional Services (0.83), Energy (0.83), and Telecommunication Services (0.71), while maintaining 1.00 precision in the remaining industry groups.

Table 4: F1-Score Comparison Table (All Classes × Models)

Industry Group	LR	RFC	SVC
Automobiles & Components	1.00	1.00	0.91
Banks	1.00	1.00	1.00
Capital Goods	1.00	1.00	0.83
Commercial & Professional Services	1.00	1.00	0.91
Consumer Discretionary Dist. & Retail	1.00	0.89	0.89
Consumer Durables & Apparel	1.00	1.00	1.00
Consumer Services	1.00	1.00	0.89
Consumer Staples Dist. & Retail	1.00	1.00	1.00
Energy	1.00	1.00	0.91
Equity REITs	1.00	1.00	1.00
Financial Services	1.00	1.00	1.00
Food, Beverage & Tobacco	1.00	1.00	1.00
Health Care Equip. & Services	1.00	1.00	1.00
Household & Personal Products	1.00	1.00	1.00
Insurance	1.00	1.00	1.00
Materials	1.00	0.91	0.57
Media & Entertainment	1.00	1.00	0.75
Pharma, Biotech & Life Sciences	1.00	1.00	1.00
Real Estate Mgmt. & Dev.	1.00	1.00	1.00
Semiconductors & Equip.	1.00	1.00	1.00
Software & Services	1.00	1.00	1.00
Technology Hardware & Equip.	1.00	1.00	1.00
Telecommunication Services	1.00	1.00	0.83
Transportation	1.00	1.00	1.00
Utilities	1.00	1.00	1.00

Table 4 reports the harmonic mean of precision and recall for each class. The LR model again demonstrates uniform performance, achieving an F1-score of 1.00 across all industry groups. The RFC model shows strong and stable results, with most classes scoring 1.00, and minor reductions observed in Consumer Discretionary Distribution & Retail (0.89) and Materials (0.91). The SVC model displays greater variability, with reduced F1-scores in Materials (0.57), Media & Entertainment (0.75), Capital

Goods (0.83), Telecommunication Services (0.83), and Consumer Discretionary Distribution & Retail (0.89), highlighting comparatively weaker class-level balance between precision and recall.

5.CONCLUSION

The experimental results and system demonstrations clearly validate the effectiveness of the proposed ROBERTA-TAXONOMY framework for automated Industry Group classification. Across 25 industry groups and 125 test samples, the DistilRoBERTa + Logistic Regression model achieved perfect performance with 100.00% accuracy, precision, recall, and F1-score, indicating complete alignment between predicted and actual industry groups. The Random Forest classifier also demonstrated near-optimal performance, achieving 99.20% accuracy, 99.33% precision, 99.20% recall, and a 99.19% F1-score, with only minor recall degradation in select consumer-related categories. In comparison, the Support Vector Classifier achieved 94.40% accuracy and an F1-score of 93.97%, showing comparatively lower robustness, particularly for complex and semantically overlapping classes such as Materials and Media & Entertainment. These quantitative results confirm that transformer-based embeddings significantly enhance class separability when combined with appropriate classifiers.

Furthermore, the consistency observed between Jupyter-based predictions and web application outputs confirms the reliability and stability of the deployed model. The prediction results for Energy-sector sub-industries, such as *Oil & Gas Drilling* and *Integrated Oil & Gas*, were correctly classified under the *Energy* industry group in both environments, demonstrating end-to-end system integrity. The intuitive web interface, combined with real-time inference, enables practical adoption of the model for automated industry taxonomy mapping. Overall, the system successfully addresses limitations of manual and rule-based classification approaches by delivering a scalable, accurate, and production-ready solution for industry group and brand-related classification tasks.

References

[1] Rizinski, Maryan, Andrej Jankov, Vignesh Sankaradas, Eugene Pinsky, Igor Mishkovski, and Dimitar Trajanov. "Comparative analysis of NLP-based models for company classification." Information 15, no. 2 (2024): 77.

[2] Angin, Merih, Beyza Taşdemir, Cenk Arda Yılmaz, Gökcan Demiralp, Mert Atay, Pelin Angin, and Gökhan Dikmener. "A roberta approach for automated processing of sustainability reports." Sustainability 14, no. 23 (2022): 16139.

[3] Areshey, Ali, and Hassan Mathkour. "Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet." Expert Systems 41, no. 11 (2024): e13701.

[4] Joshi, Hiren. "Transformer-Based Language Deep Learning Detection of Fake Reviews on Online Products." J. Electrical Systems 20, no. 3 (2024): 2368-2378.

[5] Durgam, Dr Revathi, Narendra Babu Pamula, D. R. Nallani Dharani, Siva Kumar, Dr Putta Durga, And Velchuri Balaji. "Ai Powered Empathy: Sentiment Analysis In Personal Care Using Roberta And Xlnet." Journal of Theoretical and Applied Information Technology 103, no. 8 (2025).

[6] Raza, Mubashar, Zarmina Jahangir, Muhammad Bilal Riaz, Muhammad Jasim Saeed, and Muhammad Awais Sattar. "Industrial applications of large language models." Scientific Reports 15, no. 1 (2025): 13755.

[7] Polam, Ram Mohan, Bhavana Kamarthapu, Ajay Babu Kakani, Sri Krishna Kireeti Nandiraju, Sandeep Kumar Chundru, and Srikanth Reddy Vangala. "Big Text Data Analysis for Sentiment Classification in Product Reviews Using Advanced Large Language Models." International Journal of AI, Big Data, Computational and Management Studies 2, no. 2 (2021): 55-65.

[8] Zhang, Hongzhi, and M. Omair Shafiq. "Survey of transformers and towards ensemble learning using transformers for natural language processing." Journal of big Data 11, no. 1 (2024): 25.

[9] Muhetaer, Munire, Xiaoyan Meng, Jing Zhu, Aixiding Aikebaier, Liyaer Zu, and Yawen Bai. "Symmetry and Asymmetry in Pre-Trained Transformer Models: A Comparative Study of TinyBERT, BERT, and RoBERTa for Chinese Educational Text Classification." Symmetry 17, no. 11 (2025): 1812.
[10] Liu, Zhenzhen, Juuso Eronen, Fumito Masui, and Michal Ptaszynski. "Chinese Tourist Motivations for Hokkaido, Japan: A Hybrid Approach Using Transformer Models and Statistical Methods." Tourism and Hospitality 6, no. 3 (2025): 133.
[11] Wagner, Margot, Callum Stephenson, Jasleen Jagayat, Anchan Kumar, Amir Shirazi, Nazanin Alavi, and Mohsen Omrani. "Using large language models as a scalable mentalstatus evaluation technique." NPP—Digital Psychiatry and Neuroscience 3, no. 1 (2025): 27.
[12] Wang, Yuanhang, Yonghua Zhou, Min Zhong, Yiduo Mei, Hamido Fujita, and Hanan Aljuaid. "A hierarchical interaction multimodal model for feature fusion based on RoBERTa-Keyword-ViT." Applied Intelligence 55, no. 13 (2025): 900.
[13] Muthusami, Rathinasamy, Kandhasamy Saritha, Kolli Srinivasa Rao, Palanisamy Sugapriya, and G. Saveetha. "Interpretable stance detection in social media via topic guided transformers." Discover Artificial Intelligence 5, no. 1 (2025): 355.
[14] Ishchenko, Roman. "Recent Advances in Machine Learning Algorithms for Candidate–Job Matching." Universal Library of Engineering Technology 2, no. 4 (2025).
[15] Fetahi, Endrit, Arsim Susuri, Mentor Hamiti, Zenun Kastrati, Ercan Canhasi, and Arta Misini. "Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI." Social Network Analysis and Mining 15, no. 1 (2025): 82.