International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

BrandSense: Machine Learning Framework for Context-Aware Brand Name Analysis and Prediction

Goutham Kunamalla[1], K. Sharmila Reddy[2*], Thummala Nirjara[3], Jannu Shashank[3], Abbineni Sindhu Priya[3], Sambari Vignesh[3]

[1]Assistant Professor, [2]Associate Professor & Head, [3]UG Student, [1,2,3]Department of Computer Science and Engineering

[1,2,3]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

*Correspondence: K. Sharmila Reddy (sharmilakreddy@gmail.com)

ABSTRACT

Establishing a strong and distinctive brand identity is essential for business success; however, generating innovative, meaningful, and market-ready brand names remains a complex, subjective, and time-intensive process. This project presents an AI-powered platform designed to automate and optimize brand name generation and evaluation by leveraging advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques. The system incorporates a robust text-processing pipeline that includes tokenization, lemmatization, and feature extraction using transformer-based embeddings such as DistilRoBERTa, enabling the capture of deep semantic relationships within textual data. To ensure high predictive performance and reliability, the platform integrates multiple classification models, including Logistic Regression (LR), Random Forest Classifier (RFC), and Support Vector Machines (SVM). These models are trained on datasets balanced using the Synthetic Minority Over-sampling Technique (SMOTE), improving classification accuracy and handling class imbalance effectively. The platform is supported by an interactive Web User Interface (WUI) that facilitates both single and batch predictions, providing users with suitability scores and category classifications for generated brand names. Additionally, integrated visualization tools enhance model interpretability, allowing users to analyze performance metrics and gain insights into decision-making processes. By combining state-of-the-art NLP methods with scalable ML pipelines, this system offers a data-driven solution that minimizes manual effort and subjectivity, ultimately enabling businesses to develop impactful and competitive brand identities in rapidly evolving market environments.

Keywords: Brand identity, Brand name generation, Natural language processing, Tokenization, Lemmatization, Feature extraction, Transformer embeddings, Semantic analysis, Artificial intelligence.

1. INTRODUCTION

In a time characterized by rapid entrepreneurial growth and continuous innovation, brand name generation has become a crucial task. Creating unique and memorable brand names is essential, as they serve as the first point of interaction between businesses and consumers. However, with a growing number of startups each year, identifying distinctive and brandable names has become increasingly difficult. Conventional methods, often centered on single-word naming strategies, are limited by trademark constraints and frequently fail to deliver innovative outcomes, thereby encouraging the exploration of more advanced approaches [1,2,3].

The rise of AI technologies, including Chat GPT, has introduced significant opportunities for brand marketers to improve customer engagement, streamline communication, boost creativity, and build strong brand identities [4] Chat GPT's AI-driven capabilities, powered by natural language processing (NLP) and machine learning, enable dynamic and personalized brand interactions. NLP facilitates human-like conversations, allowing systems to respond to queries, provide recommendations, and build stronger customer relationships. Meanwhile, machine learning analyzes consumer behavior and preferences, enabling tailored content delivery for targeted audiences illustrated in figure 1.

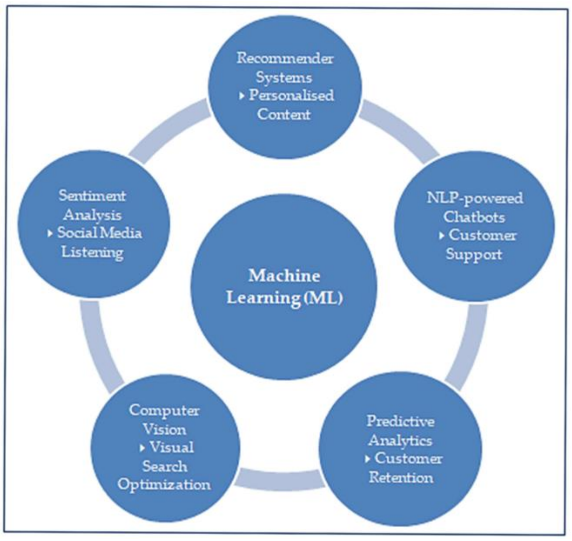


Figure 1: Context-Awareness

This integration strengthens brand communication and ensures real-time, responsive support, enhancing overall customer experience and engagement. This detailed overview explores the evolving relationship between AI and brand marketing, presenting practical insights, case studies, and recommendations for brand managers aiming to fully utilize AI tools to create impactful and customer-focused brand experiences [5]. By examining AI's role in brand communication, customer engagement, and data-driven decision-making, it offers a strategic framework for navigating the digital landscape and achieving sustained brand success in a competitive market [6].

2. LITERATURE SURVEY

Marco Lemos, et al. [7] explored techniques and datasets used to train AI models as well as strategies for leveraging objective data to validate the brand ability of generated names. Emotional and semantic aspects of brand names, which are often overlooked in traditional approaches, are discussed as well. A list with more than 75 pivotal datasets is presented. As a result, this review provided an understanding of the potential applications of AI, NLP, and affective computing in brand name generation and evaluation, offering valuable insights for entrepreneurs and researchers alike. Anuj Kumar, et al. [8] underscored the importance of responsible AI usage, transparency, and continuous adaptation to changing consumer behaviors for maintaining trust and ethical standards. This contribution to the existing literature combines real-world examples, practical insights, and a mixed-methods approach, offering a unique perspective on how AI, particularly Chat GPT, can reshape customer engagement, brand communication, and creativity in both academic and industrial contexts. The article provides a comprehensive examination of AI tools' practical utility, bridging theory and application for a nuanced understanding in the field of brand marketing.

Musaiqer, et al. [9] examined the existing literature on the use of AI in brand building, including its potential to improve customer engagement, personalize marketing strategies, and optimize brand messaging. they also highlighted the challenges associated with AI implementation, such as the need for high-quality data, potential biases in algorithms, and the risk of losing the human touch in brand interactions. Maria DSouza Deryl, et al. [10] reviewed the theoretical underpinnings of AI-driven branding from the extant literature. The thematic analysis of theoretical premises instrumentalises developing a novel integrated theoretical framework. The intricate analysis of the constellation of theories and the simultaneous amalgamation of concepts allows for a unified viewpoint. The integrated theoretical framework offers an in-depth overview of the drivers, mediators, moderators, and consequences of AI-driven branding.

Sabeh et al. [11] presented Open Brand, a model architecture designed for extracting brand values from unstructured product titles. By incorporating both word and character-level embeddings, Open Brand outperforms existing state-of-the-art models across various categories. The model demonstrates strong generalization capabilities, particularly in zero-shot extraction scenarios, and performs well even with compound brand values. Further, the problem of unseen words in brand names, which can arise due to sub-branding, brand fragmentation, or emerging businesses, is addressed. Harry Ph. Sophocleous et, al. [12] explored the intersection of these technologies with entrepreneurship, highlighting how they enhanced decision-making, customer insights, and operational efficiency. While the benefits are substantial, integrating these technologies presents challenges related to data ethics, privacy, algorithmic bias, sustainability, accessibility, and implementation complexity. Through analysis of current literature, critical perspectives, and illustrative case studies, this entry emphasizes the need for strategic alignment, ethical considerations, and adaptive organizational cultures.

Mendes et al. [13] their research evaluated the performance of multilingual DistilBERT and XLM-RoBERTa models in predicting affective ratings from textual content. The assessment took into account how model size affects the accuracy of predictions; to determine how model complexity influences the precision of valence and arousal predictions, several transformer model sizes were tested. Michael Gerlich et, al. [14] aimed to understand whether AI systems can replace human influencers in shaping purchasing decisions and, if so, in which sectors. A mixed-methods approach was employed, involving a quantitative survey of 478 participants with prior experience using both AI tools and interacting with social media influencers, complemented by 15 semi-structured interviews.

Nurmambetov et, al. [15] proposed a logistic regression model for binary sentiment analysis, starting with language identification and tokenization using tools from the Natural Language Toolkit (NLTK). They utilized NLTK’s stop-word corpus to eliminate non-informative words after tokenization, followed by lemmatization with the StanfordNLP library

3. PROPOSED SYSTEM

The proposed system employs a hybrid Natural Language Processing (NLP) and Machine Learning pipeline to classify industry groups using text data. The process begins with dataset loading and preprocessing, where raw text is cleaned, tokenized, lemmatized, and stop words are removed using NLTK as shown in the figure 2. Processed data is saved for reusability, and label encoding is applied for categorical targets. Exploratory Data Analysis (EDA) visualizations, including word clouds, POS tagging, and bigram analysis, help in understanding text structure and patterns. Semantic embeddings are extracted using the DistilRoBERTa transformer model to capture contextual relationships in textual data. The extracted embeddings are balanced using SMOTE to handle class imbalance effectively. Multiple Machine Learning algorithms, namely LRC, RFC, and SVC, are trained using these embeddings. Model performance is evaluated through metrics such as Accuracy, Precision, Recall, and F1-Score, computed by a custom Metrics Calculator module. Graphical visualizations of these metrics are generated for comparative analysis. Finally, the best-performing model is saved and used to predict industry groups for unseen test data, ensuring a robust and generalizable classification framework.

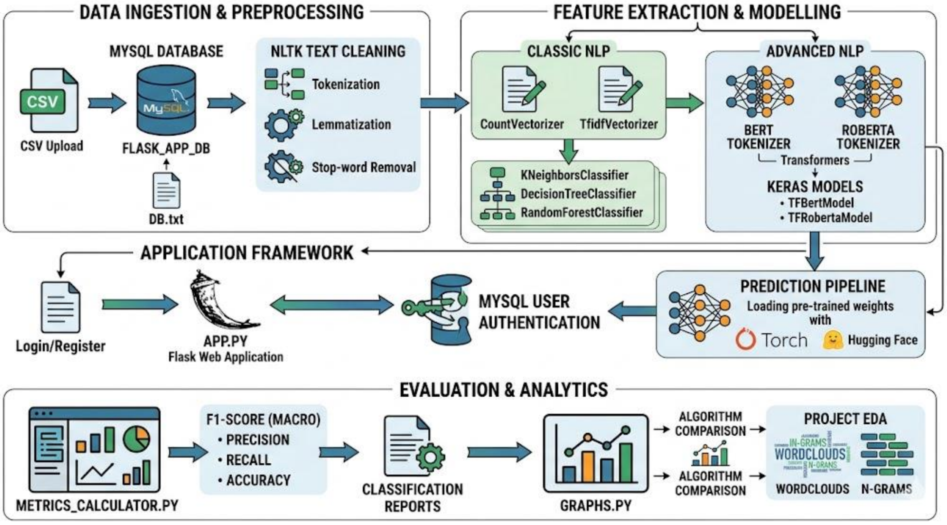


Figure 2: Proposed System Architecture

4. RESULTS ANALYSIS

The results section presents the key findings of the study in a clear and organized manner. It highlights the outcomes obtained from data analysis, experiments, or observations, often supported by tables, graphs, or figures. This section focuses only on factual information without interpretation or discussion. The results are usually aligned with the research objectives or questions, showing whether the expected outcomes were achieved. It may also include statistical data or comparisons to emphasize significant patterns or trends. It provides a concise summary of what was discovered during the research process.

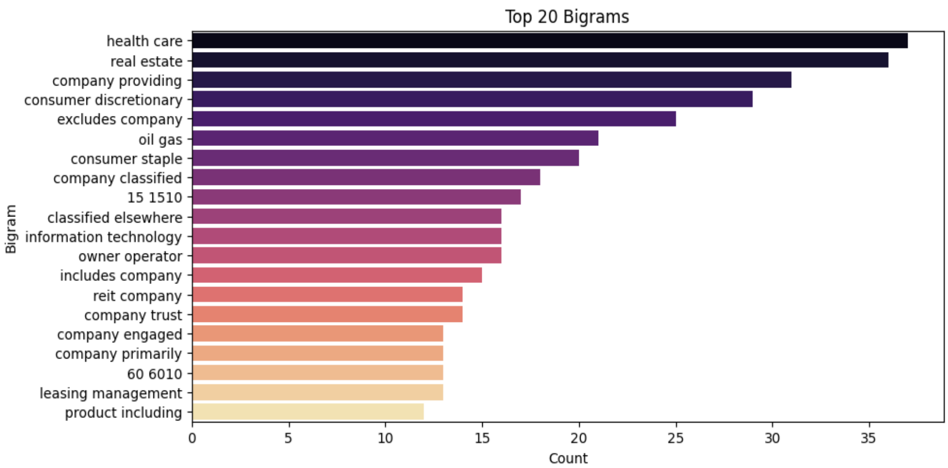
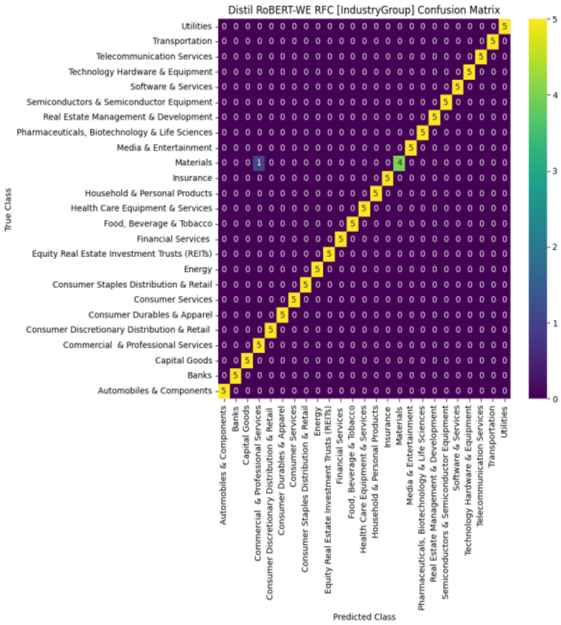


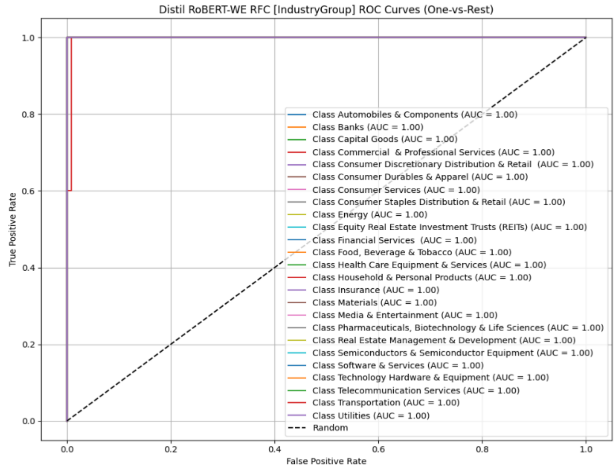
Figure 3: Top 20 Bigrams.

Figure 3 shows the top 20 most frequent bigrams extracted from the corpus, ranked by occurrence. Dominant pairs such as “health care,” “real estate,” “company providing,” and “consumer discretionary” highlight major industry domains and disclosure terminology. The repeated presence of “company” in multiple combinations reflects its central role in describing operational scope and classification. These bigrams reveal multi-word patterns that capture more specific industry semantics compared to isolated unigrams.

IJERST International Journal of Engineering Research and Science & Technology



(a)



(b)

Figure 4: Distil ROBERT-WE RFC of (a) Confusion Matrix (b) ROC Curve for Industry Group

Figure 4 (a) shows the confusion matrix for the Distil RoBERTa-WE RFC, where predictions remain consistently diagonal with only negligible isolated deviations. The near-perfect classification demonstrates that the ensemble model effectively captures non-linear relationships present in the embedding space. Minor off-diagonal entries indicate occasional confusion among semantically adjacent industry categories. Despite this, the overall structure confirms strong decision-boundary formation within the embedding-driven feature set. The RFC model exhibits excellent robustness across class variations.

Figure 4 (b) shows the ROC curves for the same model, with all industry classes achieving an AUC score of 1.00. The curves reflect ideal separation, indicating that the classifier maintains extremely low false-positive rates even for classes with subtle semantic overlap. The consistently maximal AUC scores highlight the model’s resilience and the discriminative strength of the embeddings. These curves reinforce that the RFC benefits significantly from the contextual richness encoded in transformer-based features.


Choose File No file chosen

Upload & Predict

Prediction Results

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 Fuels	10102030	Oil & Gas Refining & Marketing	Companies engaged in the refining and marketing of oil, gas and/or refined products not classified in the Integrated Oil & Gas or Independent Power Producers & Energy Traders Sub-Industries.	Energy
10	Energy	1010	101020	Oil, Gas & Consumable Fuels	10102040	Oil & Gas Storage & Transportation	Companies engaged in the storage and/or transportation of oil, gas and/or refined products. Includes diversified midstream natural gas companies, oil and refined product pipelines, coal slurry pipelines and	Energy

Figure 5: Prediction Results Page

Figure 5 shows the results interface where predicted IndustryGroup labels are displayed alongside their corresponding input attributes. The table-format presentation organizes sector, industry codes, descriptions, and model-generated outputs into a structured view. This arrangement enables users to verify predictions against actual descriptions and interpret model behavior across multiple entries. The interface supports seamless scanning of numerous records, making large-scale validation efficient and transparent. It completes the prediction pipeline by translating backend model outputs into an accessible, user-facing representation.

Table 1: Overall Performance Comparison of 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 presents the overall performance comparison of three classification models trained on Distil RoBERTa word embeddings for Industry Group prediction. The LR model achieves a perfect score of 100% across accuracy, precision, recall, and F1-score, demonstrating that the Distil RoBERTa embedding space is highly linearly separable for this task. The RFC follows closely with strong performance achieving 99.20% accuracy and similarly high precision, recall, and F1-scores indicating excellent non-linear decision boundary learning. In contrast, the SVC records comparatively lower performance, with 94.40% accuracy and an F1-score of 93.97%, revealing slight difficulty in distinguishing semantically similar industry descriptions. These results, computed on the balanced test set after SMOTE oversampling, confirm LR as the most effective classifier for IndustryGroup categorization when paired with transformer-based embeddin

5. CONCLUSION

The developed industry group classification system demonstrates the strong potential of combining transformer-based embeddings with classical machine-learning techniques for high-accuracy text classification. By leveraging DistilRoBERTa for semantic feature extraction, the model captures deep contextual meaning from industry descriptions, enabling more reliable and fine-grained predictions. Rigorous preprocessing, including tokenization, lemmatization, and data normalization, plays a crucial role in preparing high-quality textual inputs. The integration of SMOTE ensures balanced learning across all classes, reducing model bias and improving generalization. Experiments with LR, RFC, and SVC highlight that even lightweight models can achieve notable performance when powered by transformer embeddings. The system's architecture also includes comprehensive EDA, performance visualization, and caching mechanisms to enhance interpretability and efficiency. The Flask web interface offers a seamless user experience, allowing real-time predictions and streamlined industry classification workflows. Overall, the project establishes a robust, scalable, and efficient pipeline capable of handling large text-driven classification problems. The performance improvements observed across all stages confirm the effectiveness of the approach and its readiness for further expansion and real-world adoption.

REFERENCES

[1] Arora, S.; Kalro, A.D.; Sharma, D. A comprehensive framework of brand name classification. J. Brand Manag. 2015, 22, 79–116.

[2] Moro Visconti, R. Domain Name Valuation: Internet Traffic Monetization and IT Portfolio Bundling. 2017. Available online: https://ssrn.com/abstract=3028534 (accessed on 14 June 2024).

[3] Eskiev, M. Naming as one of the most important elements of brand management. SHS Web Conf. 2021, 128, 01028.

[4] Patel, S.B.; Lam, K.; Liebrenz, M. ChatGPT: Friend or foe. Lancet Digit. Health 2023, 5, e102.

[5] Doshi, R.H.; Bajaj, S.S.; Krumholz, H.M. ChatGPT: Temptations of progress. Am. J. Bioeth. 2023, 23, 6–8.

[6] George, A.S.; George, A.H. A review of ChatGPT AI's impact on several business sectors. Partn. Univers. Int. Innov. J. 2023, 1, 9–23.

[7] Lemos, M.; Cardoso, P.J.S.; Rodrigues, J.M.F. Harnessing AI and NLP Tools for Innovating Brand Name Generation and Evaluation: A Comprehensive Review. Multimodal Technol. Interact. 2024, 8, 56. https://doi.org/10.3390/mti8070056

[8] Kumar, A.; Bapat, G.; Kumar, A.; Hota, S.L.; Abishek, G.D.; Vaz, S. Unlocking Brand Excellence: Harnessing AI Tools for Enhanced Customer Engagement and Innovation. Eng. Proc. 2023, 59, 204. https://doi.org/10.3390/engproc2023059204

[9] Musaiqer, H.H., Hamdan, A. (2023). The Role of Artificial Intelligence in Brand Building: A Review. In: El Khoury, R., Nasrallah, N. (eds) Emerging Trends and Innovation in Business and Finance. Contributions to Management Science. Springer, Singapore. https://doi.org/10.1007/978-981-99-6101-6_22

[10] Maria DSouza Deryl, Sanjeev Verma, Vartika Srivastava,How does AI drive branding? Towards an integrated theoretical framework for AI-driven branding,International Journal of Information Management Data Insights, Volume 3, Issue 2,2023,100205, ISSN 2667-0968, https://doi.org/10.1016/j.jjimei.2023.100205.

[11] Sabeh, K.; Kacimi, M.; Gamper, J. OpenBrand: Open Brand Value Extraction from Product Descriptions. In Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5), Dublin, Ireland, 26 May 2022; Malmasi, S., Rokhlenko, O., Ueffing, N., Guy, I., Agichtein, E., Kallumadi, S., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 161–170.

[12] Sophocleous, H.P. Harnessing Big Data and Artificial Intelligence for Entrepreneurial Innovation: Opportunities, Challenges, and Strategic Implications. Encyclopedia 2025, 5, 122. https://doi.org/10.3390/encyclopedia5030122

[13] Mendes, G.A.; Martins, B. Quantifying valence and arousal in text with multilingual pre-trained transformers. In Proceedings of the European Conference on Information Retrieval, Dublin, Ireland, 2–6 April 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 84–100.

[14] Gerlich, M. The Shifting Influence: Comparing AI Tools and Human Influencers in Consumer Decision-Making. AI 2025, 6, 11. https://doi.org/10.3390/ai6010011

[15] Nurmambetov, D.; Dauylov, S.; Bogdanchikov, A. Kazakh Names Generator Using Deep Learning. Her. Kazakh-Br. Tech. Univ. 2021, 17, 171–177.