An AI-Driven Approach to Intelligent Naming System Design and Market-Relevance Evaluation

M. Rakesh[1*], Nune Neha[2], Chitimilla Pavan venkat[2], Bandaram Pavan kalyan[2]

[1]Assistant Professor, [2]UG Student,[1,2]Department of Computer Science and Engineering

[1,2]Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

*Correspondence: M. Rakesh (machrak3149@gmail.com)

To Cite this Article

M. Rakesh, Nune Neha, Chitimilla Pavan venkat, Bandaram Pavan kalyan, "An AI-Driven Approach to Intelligent Naming System Design and Market-Relevance Evaluation", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04, April 2026,pp: 879-887, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04.pp879-887

Submitted: 08-03-2026 Accepted: 16-04-2026 Published: 23-04-2026

Abstract

Building a compelling brand identity is essential for achieving competitive advantage, yet producing unique, meaningful, and market-aligned brand names remains a difficult and time-intensive process. Traditional techniques rely heavily on human creativity, brainstorming practices, and predefined heuristics, which are often subjective, inconsistent, and incapable of capturing deeper semantic relationships or emerging market patterns. This study presents an AI-driven system that integrates natural language processing (NLP) and machine learning to automate and enhance both the generation and evaluation of brand names. The methodology begins with preprocessing textual data through cleaning, tokenization, lemmatization, and removal of stop words to ensure high-quality input. It then utilizes transformer-based embeddings such as DistilRoBERTa to extract rich contextual and semantic representations of words and phrases. For classification tasks, multiple algorithms including Logistic Regression (LR), Random Forest Classifier (RFC), and Support Vector Machines (SVM) are implemented and trained on balanced datasets, where Synthetic Minority Over-sampling Technique (SMOTE) is applied to address class imbalance issues. The system supports both single and batch processing modes, delivering outputs such as predicted brand categories, relevance scores, and performance metrics through an interactive web interface. By combining advanced NLP techniques with robust machine learning models, the proposed platform improves efficiency, scalability, and accuracy while significantly reducing manual effort, bias, and time in developing impactful and market-ready brand names.

Keywords: Brand name generation, Natural Language Processing (NLP), transformer embeddings, DistilRoBERTa, semantic analysis, brand identity

1. Introduction

In the contemporary business ecosystem, driven by rapid entrepreneurial expansion and continuous innovation, the task of developing a compelling and distinctive brand name has become both strategically important and increasingly complex. A brand name serves as a critical touchpoint that shapes first impressions, influences consumer perception, and contributes to long-term brand recall and positioning. However, the exponential rise in startups and digital businesses has intensified

competition, making it difficult to identify names that are not only creative but also legally available and market-relevant. Conventional naming approaches, which often depend on manual brainstorming or simple linguistic constructs such as single-word names, are constrained by issues including trademark conflicts, limited originality, and lack of contextual relevance, thereby necessitating the exploration of more advanced, scalable, and intelligent methodologies [1,2,3]. With the evolution of artificial intelligence, a new paradigm has emerged that is transforming how branding strategies are conceptualized and executed. Tools such as ChatGPT have introduced significant advancements by enabling automated content generation, enhanced creativity, and more efficient communication workflows [4]. By leveraging natural language processing (NLP) and machine learning techniques, these systems can understand linguistic context, generate human-like responses, and analyze large volumes of consumer data to identify patterns and preferences. This enables businesses to deliver highly personalized and context-aware brand interactions, improving customer engagement and satisfaction. Furthermore, AI-driven solutions extend beyond communication to support critical functions such as brand name generation, semantic evaluation, sentiment analysis, and performance prediction. These capabilities empower organizations to make data-driven decisions, refine their branding strategies, and adapt quickly to changing market dynamics. Existing studies emphasize that the integration of AI into branding not only enhances efficiency and innovation but also provides a structured framework for developing impactful, customer-centric brand identities in an increasingly competitive and digitally evolving marketplace [5,6].

2. Literature Survey

Talha Ahmed Khan, et al. [7] explored real-world implementations of AI in industrial and business environments through detailed case studies. Their research emphasized the tangible benefits of AI adoption, including improved operational efficiency, enhanced decision-making, and optimized resource utilization. At the same time, they addressed practical challenges such as integration complexity, data security concerns, and ethical implications associated with AI deployment. By analyzing both successes and limitations, their work provided a balanced understanding of how organizations can effectively incorporate AI technologies while mitigating associated risks. Jiang, et al. [8] proposed a hybrid deep learning architecture that combines Word2vec-based feature extraction with an LSTM network for improved text representation and prediction. Their model captured contextual semantics by converting textual data into dense vector representations, which were then processed through the LSTM layer to learn sequential dependencies. In addition to the hybrid model, they developed several benchmark models using different ML techniques such as Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Hash Trick (HT). Experimental results showed that the Word2vec–LSTM hybrid model consistently outperformed traditional approaches in terms of prediction accuracy and robustness.

Wang, et al. [9] presented a semantic-based approach for detecting duplicate or highly similar textual content using advanced NLP techniques. Their framework integrated Word2vec embeddings with Latent Dirichlet Allocation (LDA) to capture both semantic similarity and topic-level relationships between texts. This dual-layer analysis enabled more accurate identification of paraphrased or contextually similar content compared to traditional keyword-based methods. Their work demonstrated the effectiveness of combining vector-based representations with probabilistic topic modeling for tasks involving semantic comparison and content validation. Liu, et al. [10] proposed an advanced paraphrase generation framework that extended the conventional sequence-to-sequence (S2S) architecture by incorporating an attention mechanism along with topic-aware guidance. Their model utilized topic words as prior knowledge to improve semantic consistency and contextual relevance in generated paraphrases. These topic words were extracted using Latent Dirichlet

Allocation (LDA), enabling the system to maintain thematic alignment while rewriting sentences. By integrating attention with topic modeling, the approach enhanced both fluency and diversity in paraphrase generation, addressing limitations of earlier models that often produced generic or less meaningful outputs. The study demonstrated that incorporating semantic guidance significantly improves the quality and applicability of automated text rewriting systems.

Sadhuram, et al. [11] developed an AI-driven question answering (QA) system designed to improve information retrieval and response accuracy through a structured NLP pipeline. Their implementation included multiple stages such as question analysis, passage retrieval, sentence ranking, answer extraction, and summarization. The system leveraged NLP techniques to interpret user queries effectively and identify the most relevant information from large textual datasets. It was evaluated using over 422 articles from the SQUAD dataset along with approximately 87,599 cross-domain questions, demonstrating strong performance in handling diverse query types. Their work highlighted the effectiveness of combining AI and NLP techniques to build scalable and efficient QA systems capable of delivering precise and context-aware responses. Harry Ph. Sophocleous, et al. [12] investigated the intersection of AI technologies with entrepreneurial practices, focusing on their impact on decision-making, customer insights, and operational efficiency. Their research highlighted how AI-driven analytics can support strategic planning and enhance business outcomes by providing data-driven insights. At the same time, they critically examined the challenges associated with adopting such technologies, including ethical concerns, data privacy issues, algorithmic bias, sustainability considerations, and implementation complexity. Through an analysis of existing studies and case examples, they emphasized the importance of aligning technological adoption with organizational values and fostering adaptive cultures to fully realize the benefits of AI in business environments.

Mendes, et al. [13] investigated the effectiveness of multilingual transformer-based models in predicting affective characteristics from textual data, with particular focus on dimensions such as valence and arousal. Their study compared models like DistilBERT and XLM-RoBERTa across different configurations to understand how architectural complexity and model size influence predictive performance. By experimenting with multiple model variants, they demonstrated that larger and more expressive models tend to capture emotional nuances more accurately, although at the cost of increased computational requirements. Their findings provided valuable insights into the trade-offs between efficiency and accuracy, particularly for applications requiring semantic and emotional interpretation of language. Michael Gerlich, et al. [14] examined the potential of AI systems to influence consumer behavior in comparison to human social media influencers. Using a mixed-methods approach, their research combined quantitative survey data from 478 participants with qualitative insights from semi-structured interviews. The study explored user perceptions, trust levels, and effectiveness of AI-generated recommendations across different sectors. Their results indicated that while AI systems can effectively support decision-making in certain contexts, human influencers still hold an advantage in areas requiring emotional connection and authenticity. This work highlighted the evolving role of AI in marketing and its limitations in replicating human-driven influence.

Nurmambetov, et al. [15] proposed a machine learning-based framework for binary sentiment classification using Logistic Regression (LR). Their approach incorporated a structured preprocessing pipeline, beginning with language detection and tokenization using tools from the Natural Language Toolkit (NLTK). They further refined the textual data by removing stop words and applying lemmatization with Stanford to standardize word forms. This preprocessing significantly improved feature quality and model performance. Their study demonstrated that even relatively simple models

like LR can achieve effective results when combined with robust preprocessing and feature engineering techniques.

3. Proposed Methodology

The proposed system is designed as a modular end-to-end pipeline that processes textual industry data from acquisition to deployment. It begins with loading a structured CSV dataset, followed by preprocessing steps such as tokenization, stop word removal, lemmatization, and label encoding to standardize inputs. The cleaned text is transformed into contextual embeddings using DistilRoBERTa with mean pooling, capturing deep semantic relationships. These embeddings are cached to improve efficiency and reused during training and inference. To address class imbalance, SMOTE is applied before splitting the dataset into training and testing subsets. Multiple models, including LRC, RFC and SVC are trained in parallel to learn diverse patterns. As illustrated in Figure 1, the system enables comparative evaluation of models to identify the best performer. Performance metrics such as accuracy, confusion matrix, and ROC curves are computed and visualized for analysis. Finally, the selected model is deployed using a Flask-based application, where users can upload new data, receive predictions, and store results, ensuring a complete automated workflow.

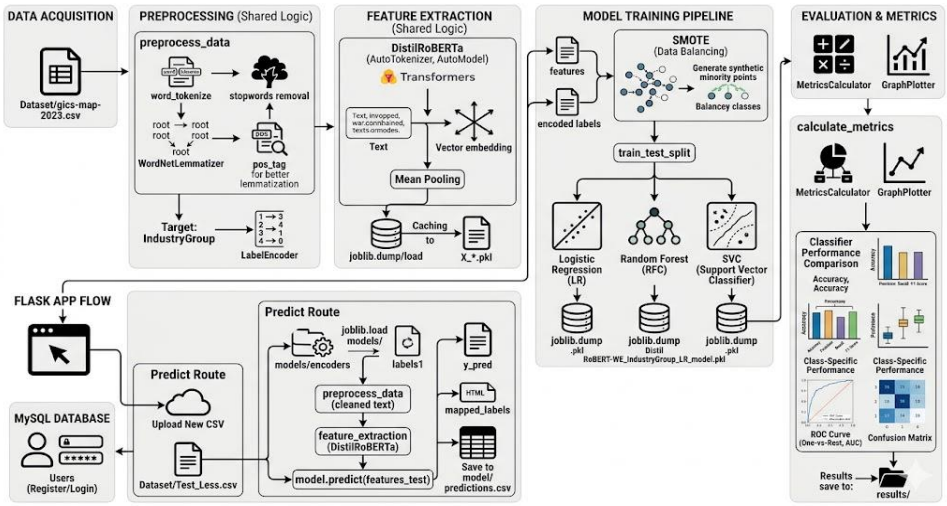

Figure. 1: Proposed System Architecture

3.1 DistilRoBERTa

DistilRoBERTa, is a lightweight and highly efficient version of RoBERTa created through knowledge distillation. It retains nearly 95% of RoBERTa language understanding capability while being significantly faster and smaller, making it ideal for real-time applications. Built on transformer self-attention, it captures deep semantic meaning and contextual relationships across text. Its compact architecture ensures quick inference without compromising on accuracy. Because of this balance between speed and performance, DistilRoBERTa is widely used for text encoding, classification, semantic analysis, and large-scale NLP deployments.

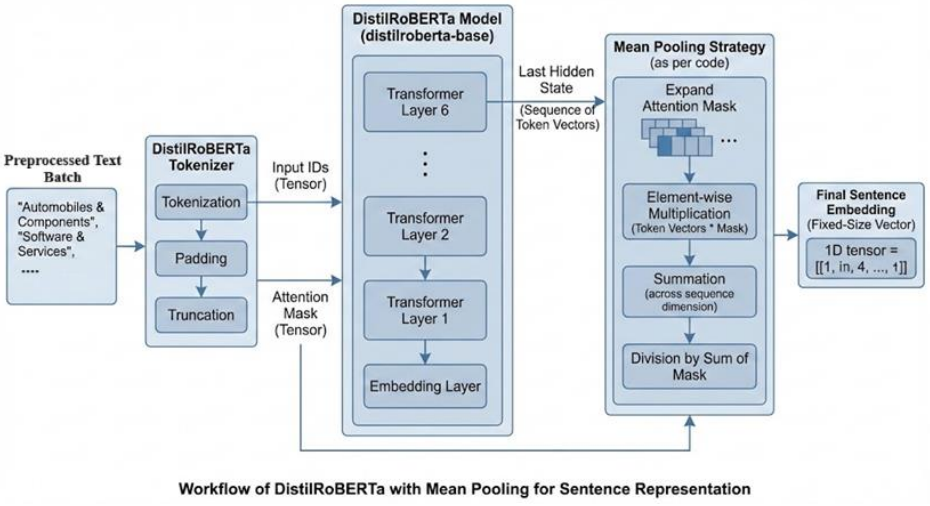

Figure. 2: Internal working flow of DistilROBERTa features extraction.

In this research, DistilRoBERTa is used as a feature extractor to convert each cleaned text entry into a dense semantic embedding vector. The code loads the pretrained "distil Roberta-base" model and tokenizer from Hugging Face, tokenizes each text, applies padding/truncation, and passes it through the transformer, as shown in figure 2. Instead of using the CLS token, the pipeline applies mean pooling across token embeddings to obtain a context-rich sentence representation. These embeddings are then saved and fed into machine learning classifiers (LR, RFC, SVC) for predicting the Industry Group labels. By using DistilRoBERTa this way, the system gains powerful contextual understanding while keeping processing fast and efficient for a live Flask application.

Pre-processed Text Input: The cleaned text generated from the NLP preprocessing stage is collected into a unified list of processed sentences. This ensures all noise, stop words, and unnecessary symbols are removed before transformer encoding.

Tokenization + Attention Mask Creation (DistilRoBERTa Tokenizer): Each processed sentence is converted into token IDs while generating an attention mask that marks real tokens vs. padding. Padding and truncation ensure all sequences follow a consistent length for batch processing.

Embedding Layer Conversion: The token IDs are passed into the embedding layer of DistilRoBERTa, where each token is mapped to a high-dimensional vector. These embeddings serve as the initial semantic representation for the model.

Transformer Encoder Processing (Layers 1–6): The embeddings flow through six transformer layers where self-attention and feed-forward networks refine token-level representations. Each layer adds deeper contextual understanding to every token vector.

Extraction of Last Hidden States: DistilRoBERTa outputs a sequence of hidden states representing each token after the final transformer layer. These vectors contain the rich semantic information needed for classification.

Mean Pooling for Sentence Representation: The hidden states are multiplied by the attention mask, summed across the sequence, and divided by the number of valid tokens. This produces a single fixed-size embedding that represents the entire sentence.

Feature Matrix Formation for Balancing: All sentence embeddings are stacked to form the complete feature matrix. These feature vectors are now ready for the next stage of the pipeline SMOTE balancing.

SMOTE Balancing on Feature Embeddings: SMOTE oversamples minority classes using the feature vectors, ensuring all Industry Group categories have equal representation. This balanced dataset is then used for ML model training.

4. Results and Description

Result analysis is a crucial stage in any study or experiment, as it involves interpreting the collected data to draw meaningful conclusions. It helps in identifying patterns, relationships, and trends within the results. Through analysis, raw data is transformed into useful information that supports or refutes the research objectives. This process also allows researchers to evaluate the accuracy and reliability of their findings. By comparing expected outcomes with actual results, one can determine the success or limitations of the study. Additionally, result analysis aids in making informed decisions and recommendations for future research or practical applications.

Figure 3 shows the confusion matrix for the Distil RoBERTa-WE-SVC on Industry Group classification. The matrix predominantly follows a perfect diagonal structure, with a few sparse misclassifications concentrated in classes with overlapping business terminology. These errors reflect the strict margin-based decision boundaries characteristic of SVCs, particularly when handling closely related categories. Even with these isolated deviations, overall classification remains highly accurate due to the dense and meaningful representation learned by the embeddings. The visualization confirms SVC's strong performance in high-dimensional semantic spaces.

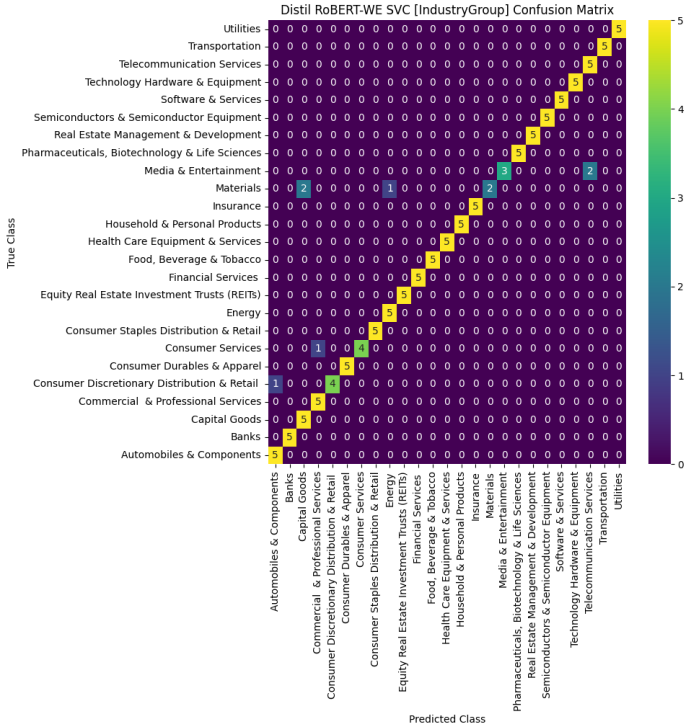

Figure. 3: Distil ROBERT-WE-SVC of Confusion Matrix for Industry Group

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. 4: Prediction Results Page

Figure 4 shows the results interface where predicted Industry Group 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 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 Industry Group categorization when paired with transformer-based embeddings.

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

5. Conclusion

The proposed industry classification framework effectively demonstrates the capability of integrating transformer-based representations with traditional machine learning models for accurate text

categorization. By utilizing DistilRoBERTa embeddings, the system captures rich contextual semantics from textual descriptions, leading to more precise predictions. Careful preprocessing steps such as tokenization, lemmatization, and normalization significantly enhance input data quality. The application of SMOTE helps in addressing class imbalance, ensuring fair learning across all categories. Models like LR, RFC, and SVC show strong performance when combined with these deep semantic features. The pipeline also incorporates detailed EDA, visualization techniques, and efficient caching to improve analysis and speed. A Flask-based interface enables smooth user interaction with real-time prediction capabilities. The solution presents a scalable and reliable approach suitable for practical deployment in large-scale text classification tasks.

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

[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] Khan, T.A.; Ali, S.M.; Ali, K.M.; Aziz, A.; Ahmad, S.; Anwar, A.; Khan, S.A. Harnessing Artificial Intelligence for Optimum Performance in Industrial Automation. Eng. Proc. 2024, 76, 105. https://doi.org/10.3390/engproc2024076105

[8] Jiang, H.; Hu, C.; Jiang, F. Text Sentiment Analysis of Movie Reviews Based on Word2Vec-LSTM. In Proceedings of the 14th International Conference on Advanced Computational Intelligence (ICACI), Wuhan, China, 22 July 2022; pp. 129–134.

[9] Wang, X.; Dong, X.; Chen, S. Text duplicated-checking algorithm implementation based on natural language semantic analysis. In Proceedings of the IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 5 June 2020; pp. 732–735.

[10] Liu, Y.; Lin, Z.; Liu, F.; Dai, Q.; Wang, W. Generating paraphrase with topic as prior knowledge. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 2381–2384.

[11] Sadhuram, M.V.; Soni, A. Natural language processing based new approach to design factoid question answering system. In Proceedings of the 2nd International Conference on Inventive Research in Computing Applications (ICIRCA), Virtual, 15–17 July 2020; pp. 276–281

[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.