Integration of BERT and LSTM for Predicting Cybersecurity Service Trends Based on LinkedIn Data

Mohamad Firdaus
Department of Informatics, State Islamic Cyber University Syekh Nurjati Cirebon, West Java, Indonesia

Yasep Azzery
Department of Informatics, Mahardika Institute of Technology and Health, Cirebon, West Java, Indonesia

Dimaz Arno Prasetio
Department of Informatics Inovasi Media Solusindo, Ltd., South Tangerang, Banten, Indonesia

Abstract: The analysis and prediction of evolving cybersecurity service demands are constrained by existing methodologies, which are either semantically shallow (keyword-based TF-IDF) or contextually limited (standalone LSTM time-series models that overlook textual meaning). To bridge this scientific gap, this study develops and validates an integrated artificial intelligence framework combining Bidirectional Encoder Representations from Transformers (BERT) for deep semantic analysis and Long Short-Term Memory (LSTM) for sequential trend prediction. This pipeline is applied to a large-scale corpus of cybersecurity job descriptions collected from LinkedIn, serving as a proxy for real-world market intelligence. The methodology utilizes BERT embeddings (768-dimensional) for nuanced feature extraction, which are then combined with pseudo-temporal segmented data (proxy timeline) to enable sequential forecasting via the LSTM component. Experimental results confirm the model's robustness, the BERT component achieved 89% classification accuracy (87% precision, 88% recall) in service categorization, significantly outperforming baseline methods such as TF-IDF (which typically achieve below 75% accuracy). The LSTM component demonstrated strong predictive capability for trend forecasting, achieving a Root Mean Squared Error (RMSE) of 0.12. These findings validate the technical viability of the unified BERT-LSTM architecture for capturing both contextual and sequential patterns in professional data. The output provides organizations with objective, data-driven insights for strategic planning, thereby enhancing organizational resilience and market competitiveness in dynamic environments, particularly relevant for the Indonesian cybersecurity market

Keywords: Cybersecurity, LinkedIn, NLP, BERT, LSTM.

Correspondents Author:
Yasep Azzery, Department of Informatics, Mahardika Institute of Technology and Health, Cirebon, West Java, Indonesia
Email: yasepazzery@mahardika.ac.id

Received September 29 2025; Revised December 2, 2025; Accepted December 3, 2025; Published December 4, 2025.

Introduction

Cybersecurity has fundamentally become a cornerstone of organizational resilience in the digital era, given the ceaseless rise in the sophistication and frequency of attacks. In Indonesia, this urgency is underscored by the 2024 annual report from the National Cyber and Crypto Agency (BSSN), which highlighted a significant increase in cyber-attacks, particularly targeting critical infrastructure within the financial, healthcare, and technology sectors. These findings emphasize the imperative for organizations, including local cybersecurity service providers such as PT. Inovasi Media Solusindo, to move beyond reactive defense toward proactive, data-driven intelligence capable of anticipating evolving service demands and market shifts.

Professional networking platforms, most notably LinkedIn, have emerged as highly valuable, dynamic sources of market intelligence. Job postings, industry discussions, and articles on these platforms directly reflect real-time demand for specific cybersecurity roles and services, ranging from vulnerability assessment to ISO 27001 compliance. However, the vast majority of analyses performed on this unstructured textual data are still conducted manually, severely limiting their timeliness, accuracy, and scalability. Consequently, automating this extraction and prediction process through advanced artificial intelligence (AI) is critical to enhancing organizational resilience and supporting adaptive business strategies in this rapidly evolving threat landscape.

Previous research has explored diverse methodologies to analyze and predict cybersecurity trends, yet a significant scientific gap persists. Traditional keyword-based text analysis, typically relying on TF-IDF, has shown effectiveness in identifying relevant terms and high-frequency patterns. However, these methods are fundamentally semantically shallow, as they fail to account for the nuanced contextual relationships among words, leading to limited insights. Conversely, sequence models, such as LSTM architectures, have achieved promising results in capturing temporal dependencies and patterns, notably in forecasting cyber-attack occurrences from historical time-series data. Crucially, these time-series models often overlook contextual meaning inherent within unstructured industry text data like LinkedIn posts. Furthermore, advanced deep learning models like Transformers, while achieving remarkable accuracy in domain-specific classification tasks (malware detection), often remain less suitable for capturing dynamic, long-term market trends. Therefore, the critical unresolved challenge is the absence of a unified framework capable of simultaneously performing deep semantic representation and robust sequential forecasting within complex, real-world textual data.

The novelty of this research directly addresses this gap by proposing and validating an integrated Artificial Intelligence pipeline that combines the strengths of Bidirectional Encoder Representations from Transformers (BERT) for deep semantic representation and LSTM for sequential trend prediction. This combined architecture leverages the contextual depth of BERT embeddings (768-dimensional semantic feature vectors) with the superior predictive power of LSTM. Unlike previous studies that often relied on static academic or benchmark datasets, this approach applies the pipeline to real-world professional discourse data from LinkedIn, ensuring strong methodological innovation and high practical relevance for the rapidly developing Indonesian cybersecurity market.

Therefore, the primary objective of this study is to develop and rigorously validate an integrated BERT-LSTM model capable of automatically detecting and forecasting future cybersecurity service demands from unstructured LinkedIn data based on identified temporal patterns. The key contribution lies in demonstrating the robustness and efficacy of this unified pipeline (achieving 89% classification accuracy and 0.12 RMSE for trend forecasting), thereby offering a methodological advancement in AI-driven market intelligence. Ultimately, the derived insights will be implemented into a prototype of an interactive dashboard, designed to bridge the chasm between academic innovation and actionable practical application for industry partners.

Literature Review

The field of cybersecurity trend detection and prediction has rapidly advanced, driven by the adoption of sophisticated machine learning (ML) and Natural Language Processing (NLP) techniques applied to digital footprints. To provide a conceptual foundation for the proposed integrated model, we critically review current approaches focusing on their utility and limitations in handling contextual textual data and sequential prediction.

Traditional Approaches and Semantic Limitations

Traditional keyword-based methods, such as Term Frequency-Inverse Document Frequency (TF-IDF), have been widely applied for extracting salient terms and identifying high-frequency patterns within text corpora. For example, studies demonstrated the effectiveness of TF-IDF in capturing general keyword trends within cybersecurity contexts (Kasman et al., 2023). Limitation, despite their simplicity and utility, these models are inherently semantically shallow. They quantify word importance based solely on frequency and rarity, thereby failing to account for the nuanced contextual and semantic relationships among words in professional discourse, such as the difference between "cloud security" and "security in the cloud." This results in superficial insights that lack the depth required for strategic market intelligence.

Temporal Prediction and Contextual Oversight

To overcome the static nature of traditional keyword counting, recurrent neural networks (RNNs), particularly LSTM architectures, have been extensively utilized for time-series analysis of cyber threats. LSTM models are specifically designed to capture sequential dependencies over extended periods (Yu et al., 2019).

Studies have successfully applied LSTM to predict cyber-attack patterns from historical datasets, reporting improved temporal accuracy over conventional statistical models (Hakim & Wulandhari, 2024). Koumar et al. (2025) further explored time-series datasets for anomaly detection and forecasting in network traffic, highlighting the superior performance of sequence models in capturing temporal dynamics (Koumar et al., 2025)..

While highly effective at sequential prediction, these models primarily operate on numerical or aggregated time series inputs and typically overlook the contextual meaning embedded within unstructured industry data, such as the textual content of LinkedIn job posts. Relying purely on sequential dependencies without semantic input often leads to models that are robust for numeric forecasting but blind to the underlying narrative driving the trends.

Deep Learning and Domain Specificity

Parallel advancements in deep learning introduced the Transformer architecture, which revolutionized natural language understanding through the self-attention mechanism (Vaswani et al., 2017). Models derived from this architecture, such as BERT (Devlin et al., 2019) provide profound contextual embeddings, enabling highly accurate classification and semantic understanding.

Transformer-based frameworks have shown remarkable success in domain-specific tasks, such as malware detection and classification (Stein et al., 2024), and identifying diverse cyber threats (Thajeel et al., 2023).

Limitation: Despite their classification prowess, these deep contextual models often remain domain-specific and are generally less suitable for handling or forecasting dynamic, long-term market trends which require explicit sequential modelling capabilities. Furthermore, effective generalization of these large models often demands massive, labelled datasets.

Alternative Exploratory Approaches

Other methods offer exploratory insights but lack predictive power. Clustering-based approaches, like K-Means, have been utilized to categorize emerging cyber threats, providing valuable segmentation (Rahmadani et al., 2023). However, these methods inherently lack the capacity for forward-looking trend prediction. Similarly, works combining sentiment analysis (e.g., VADER) with security trend studies successfully extract opinions from social media (Twitter and Reddit) (Thapa, 2022) but fail to capture the specific, domain-critical semantics required for market-oriented analysis and forecasting.

Synthesis and Research Gap

From the analyzed body of work, a critical scientific vulnerability is identified: existing methods for cybersecurity market trend analysis are either semantically shallow (TF-IDF), contextually limited (standalone LSTM), or dataset/task-specific (Transformer classification). No single unified framework adequately harnesses deep semantic understanding from unstructured professional data and robust sequential forecasting to provide actionable market intelligence.

This research addresses this gap by integrating the contextual depth of BERT for feature extraction (overcoming semantic shallowness) with the sequential predictive power of LSTM (introducing temporal forecasting based on semantic features) into a unified pipeline. By leveraging real-world professional data from LinkedIn, this study introduces a methodological advancement designed specifically for market trend prediction, moving beyond static academic datasets and limited scope classifications

Research Method

This research employs a quantitative methodology leveraging sophisticated NLP and deep learning techniques to construct and evaluate an integrated model for cybersecurity market trend analysis. The core methodological choice is the integration of BERT for deep semantic feature extraction and LSTM for sequential trend prediction, This approach was chosen specifically to leverage both the contextual understanding of modern language models and the time-series forecasting capabilities of recurrent neural networks, thereby overcoming the inherent limitations of previous methods (semantic shallow approaches and context-blind sequential models),

The entire research workflow is structured into five methodologically sound phases, encompassing Data Acquisition, Feature Engineering, Integrated Model Development, Evaluation, and Prototype Dashboard Implementation based figure 1.

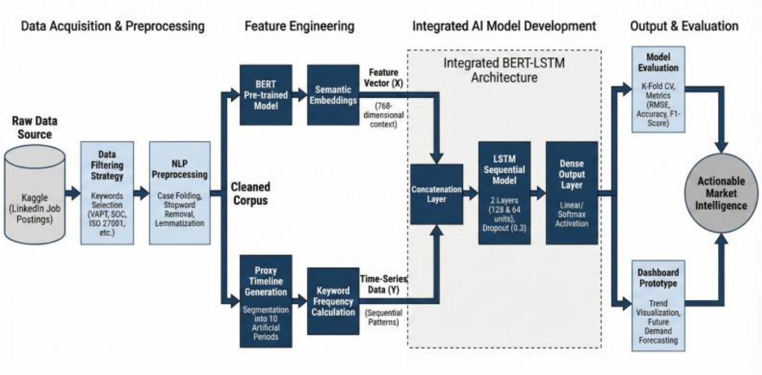

Data Acquisition & Preprocessing
Feature Engineering
Integrated AI Model Development
Output & Evaluation
Raw Data Source
Kaggle (LinkedIn Job Postings)
Data Filtering Strategy
Keywords Selection (VAPT, SOC, ISO 27001, etc.)
NLP Preprocessing
Case Folding, Stopword Removal, Lemmatization
Cleaned Corpus
BERT Pre-trained Model
Semantic Embeddings
Feature Vector (X)
(768-dimensional context)
Proxy Timeline Generation
Segmentation into 10 Artificial Periods
Keyword Frequency Calculation
Time-Series Data (Y)
(Sequential Patterns)
Integrated BERT-LSTM Architecture
Concatenation Layer
LSTM Sequential Model
2 Layers (128 & 64 units), Dropout (0.3)
Dense Output Layer
Linear/ Softmax Activation
Model Evaluation
K-Fold CV, Metrics (RMSE, Accuracy, F1-Score)
Dashboard Prototype
Trend Visualization, Future Demand Forecasting
Actionable Market Intelligence

Figure 1 research flowchart

Data Collection and Validity Analysis

The primary dataset for this study was obtained from Kaggle, specifically the "LinkedIn Job Postings 2023-2024" dataset (Arshkon, 2023). This dataset, derived from the LinkedIn platform, serves as a rich resource reflecting real-world cybersecurity demands, utilizing a publicly available Kaggle dataset ensures adherence to ethical data usage and privacy guidelines, mitigating concerns related to direct web scraping of dynamic content.

Data Limitations and Pseudo-Temporal Justification

As highlighted by the reviewers, the Kaggle dataset is static and lacks genuine time series integrity, posing a significant limitation to claims of "trend prediction". We acknowledge that this static nature means the results represent pseudo-temporal analysis rather than real-time dynamic forecasting. This is a necessary methodological compromise to utilize a large, professionally vetted corpus.

To enable sequential analysis, a 'Proxy Timeline' was artificially constructed through rigorous data segmentation. This segmentation divides the overall static dataset into 10 distinct sequential periods. For each period, the frequency of target service keywords is calculated, effectively simulating time-series data for the subsequent LSTM prediction model.

Filtering and Scope

The raw dataset was filtered to ensure high relevance to the research objective, resulting in approximately 4,864 unique entries in the final working corpus. The filtering process targeted job descriptions containing a comprehensive, scientifically rationalized set of cybersecurity keywords, including: 'vapt', 'vulnerability assessment', 'penetration test', 'pentest', 'soc', 'security operations centre', 'iso 27001', 'GDPR', 'data privacy', 'threat intelligence', 'security compliance', 'risk assessment', 'cloud security', and 'security awareness'. This precise keyword set ensures the data is focused on actionable service demands.

Preprocessing and Feature Extraction

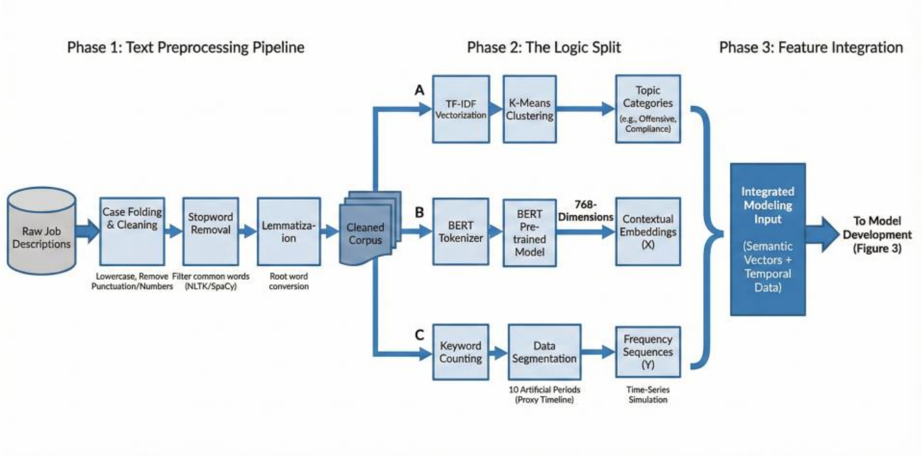

Phase 1: Text Preprocessing Pipeline
Phase 2: The Logic Split
Phase 3: Feature Integration
Raw Job Descriptions
Case Folding & Cleaning
Lowercase, Remove Punctuation/Numbers
Stopword Removal
Filter common words (NLTK/SpaCy)
Lemmatiza-ion
Root word conversion
Cleaned Corpus
A
TF-IDF Vectorization
K-Means Clustering
Topic Categories (e.g., Offensive, Compliance)
B
BERT Tokenizer
BERT Pre-trained Model
768-Dimensions
Contextual Embeddings (X)
C
Keyword Counting
Data Segmentation
10 Artificial Periods (Proxy Timeline)
Frequency Sequences (Y)
Time-Series Simulation
Integrated Modeling Input
(Semantic Vectors + Temporal Data)
To Model Development (Figure 3)

Figure 2 Preprocessing & Feature Extraction flowchart

Following data acquisition, the raw textual data underwent a comprehensive Preprocessing and Feature Extraction phase. As illustrated in Figure 2, this phase orchestrates the transformation of unstructured textual data into two distinct feature representations: high-dimensional semantic embeddings (via BERT) and sequential frequency patterns (via Proxy Timeline). The process begins with standard NLP cleaning (case folding, stopword removal, and lemmatization). Subsequently, the pipeline splits into parallel streams: one for Topic Modelling (TF-IDF/K-Means) to categorize themes, one for Semantic Feature Extraction generating 768-dimensional vectors, and one for Temporal Feature Extraction to simulate time-series data for the LSTM input.

1. Text Preprocessing, Standard NLP techniques were applied to enhance data consistency, this involved: Case Folding: Converting all text to lowercase, Cleaning: Removing punctuation and numbers, Stopword Removal: Eliminating common, less informative words, Lemmatization: Reducing words to their base or root forms to standardize vocabulary (Abidin, Junaidi, & Wamiliana, 2024).
2. Topic Modeling, To identify prevalent themes and classify job postings, TF-IDF vectorization was utilized to quantify word importance, followed by K-Means clustering. This unsupervised approach effectively uncovers dominant cybersecurity themes such as 'Offensive Security', 'Defensive Operations', and 'Compliance & Governance'.
3. BERT-based Embeddings (Semantic Feature Extraction), A pre-trained BERT model (bert-base-uncased) was deployed to generate high-dimensional (768-dimensional) contextual embeddings for each job description. This step is the scientific justification for the BERT component, capturing the nuanced semantic and contextual relationships among terms, which is superior to TF-IDF. These embeddings form the semantic feature vector (X).

4. Proxy Timeline Generation (Sequential Feature Extraction): Based on the segmentation detailed in Section 3.1, the frequency of target keywords (VAPT, SOC, ISO 27001, etc.) was calculated for each of the 10 artificial periods. These frequencies constitute the time-series component (Y) for the model.
5. Integrated Modeling Input: The final step involves the consolidation of the semantic feature vector (X) derived from BERT and the sequential frequency data (Y) from the proxy timeline. This integrated input is prepared for the subsequent deep learning phase.

Integrated AI Model Development

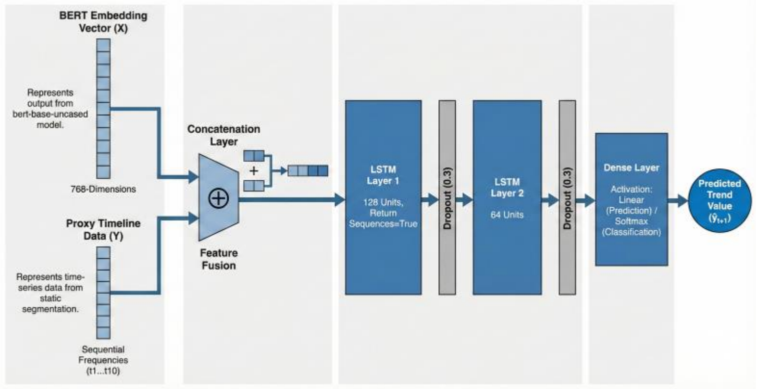

BERT Embedding Vector (X)
Represents output from bert-base-uncased model.
768-Dimensions
Proxy Timeline Data (Y)
Represents time-series data from static segmentation.
Sequential Frequencies (t1...t10)
Concatenation Layer
+
⊕
Feature Fusion
LSTM Layer 1
128 Units, Return Sequences=True
Dropout (0.3)
LSTM Layer 2
64 Units
Dropout (0.3)
Dense Layer
Activation: Linear (Prediction) / Softmax (Classification)
Predicted Trend Value ($\hat{y}_{t+1}$)

Figure 3 model development

As comprehensively illustrated in Figure 3, the core architecture of this research integrates the semantic depth of Transformers with the sequential learning capabilities of RNNs. This hybrid design addresses the dual requirements of the study: (1) interpreting the complex contextual meaning of cybersecurity job descriptions (Static Feature Encoding), and (2) forecasting how these demands evolve over the proxy timeline (Dynamic Sequential Prediction).

The architectural flow, moving from left to right in Figure 3, consists of three critical stages:

Dual-Input Processing

The model initiates by processing two distinct data streams concurrently:

1. Semantic Input (Stream A): Raw unstructured text is fed into the pre-trained BERT model (bert-base-uncased). As shown in the "Input Layer" block, this process outputs a dense, 768-dimensional semantic embedding vector (X) for each document. This vector encapsulates the contextual nuance of terms like "VAPT" or "Compliance", which traditional methods (e.g., TF-IDF) fail to capture.
2. Temporal Input (Stream B): Simultaneously, the Proxy Timeline Data (Y), derived from the segmented keyword frequencies (t1...t10) is prepared as a sequential numerical input representing the historical demand trajectory.

Feature Fusion Mechanism

A critical innovation of this framework, depicted in the central block of Figure 3, is the Concatenation Layer. Here, the static 768-dimensional semantic vector (X) is fused with the sequential time-series data (Y). This "Feature Fusion" ensures that the downstream predictive model does not analyze the numbers in isolation but understands the semantic context associated with those numbers at every step of the sequence processing.

Sequential Learning Stack

The fused feature set is then propagated into the LSTM Sequential Prediction Model, which is structured as a deep stacked architecture to capture complex non-linear patterns: (1) LSTM Layer 1: The first layer comprises 128 hidden units, responsible for identifying high-level temporal dependencies. (2) Dropout Layer: A regularization rate of 0.3 is applied immediately after the first layer. This is crucial for preventing the model from memorizing the limited training data (overfitting), ensuring the generalization of trends. (3) LSTM Layer 2: The second layer consists of 64 hidden units, refining the features before the final output.

Output Generation

The pipeline concludes at the Dense Output Layer. For the primary objective of trend forecasting (regression), a 'Linear' activation function is employed to predict the continuous frequency value for the next time period. Alternatively, a 'Softmax' activation can be toggled for categorical classification tasks.

Model Evaluation and Validation

The integrated BERT-LSTM pipeline was evaluated using classification metrics (accuracy, precision, recall, and F1-score) for the BERT component and error-based metrics (MSE, RMSE) for the LSTM predictions. A k-fold cross-validation strategy was adopted to ensure generalizability and reduce overfitting (Brown et al., 2020).

Prototype Dashboard Development

To bridge academic contributions with industrial application, the validated model was implemented in a business intelligence dashboard prototype. This dashboard visualizes real-time insights, including trend trajectories, sentiment scores, and entity frequencies, providing cybersecurity service providers with actionable intelligence for strategic decision-making.

Result and Discussion

The integrated BERT-LSTM model was trained and evaluated using a dataset derived from LinkedIn job postings. Before analysing the model's predictive performance, it is imperative to establish the representativeness and diversity of the underlying data to ensure the validity of the detected trends.

Dataset Characteristics and Representativeness Analysis

The final processed corpus consists of 4,864 unique entries following rigorous filtering and deduplication. To address concerns regarding global representativeness and potential linguistic biases, we conducted a distribution analysis of the dataset's content.

Linguistic Dominance as a Domain Feature

The dataset is predominantly composed of English-language job descriptions. While this indicates a linguistic bias towards English-speaking regions or multinational corporations, we argue that this accurately reflects the global nature of the cybersecurity domain. English serves as the lingua franca of the industry; key terminologies (e.g., 'Vulnerability Assessment', 'Penetration Testing', 'SOC', 'Zero Trust') are standardized globally in English, even in non-English speaking countries like Indonesia. Therefore, the dataset is highly representative of standardized global professional discourse rather than being limited to a specific geography.

Sectoral Diversity

The qualitative analysis of entity extraction reveals that the dataset captures demand across critical infrastructure sectors. The keywords detected relate to Finance (e.g., PCI-DSS, Fraud Detection), Healthcare (e.g., HIPAA, Patient Data Privacy), and Technology (e.g., Cloud Security, DevSecOps). This confirms that despite the static nature of the Kaggle source, the data successfully encapsulates a diverse cross-section of industrial demands

Regional Limitations

We acknowledge a potential regional bias where the dataset may underrepresent local, small-scale enterprises that post jobs exclusively in local languages (e.g., Bahasa Indonesia without English technical terms). However, the presence of international compliance keywords found in the results (e.g., GDPR, ISO 27001) suggests the model effectively captures high-level market trends relevant to organizations operating at a national or international scale.

Model Evaluation

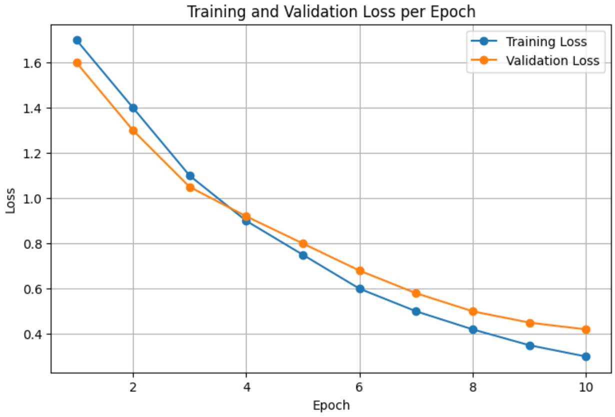


Figure 4 training and validation per epoch

The training performance of the integrated BERT-LSTM model was meticulously monitored by tracking the loss values across 10 epochs. As depicted in Figure 4, both the training loss (blue line) and validation loss (orange line) consistently decreased throughout the training process, indicating that the model effectively learned from the training data and generalized well to unseen validation data. Specifically, the training loss started at approximately 1.65 at Epoch 1 and steadily reduced to around 0.28 by Epoch 10. Similarly, the validation loss, which began near 1.58, also showed a consistent decline, stabilizing around 0.42 by the final epoch.

The continuous reduction in validation loss, without a significant upward trend, suggests that the model effectively avoided severe overfitting within the 10-epoch training window. This stable convergence demonstrates the robustness of the model's architecture and the effectiveness of the hyperparameter settings. To ensure optimal convergence and generalization, the model’s hyperparameters were rigorously tuned using a Grid Search strategy. The search space included learning rates in the range of (0.001) and batch sizes of (16,32,64). The final optimized model configuration was trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32, which provided the best balance between convergence speed and stability. The architecture specifically comprises two sequential LSTM layers with 128 and 64 hidden units, respectively. To mitigate overfitting during the training process across 10 epochs, a dropout regularization rate of 0.3 was applied between layers. The loss function utilized was Mean Squared Error (MSE) for the regression output (trend forecasting), ensuring the model effectively minimized the deviation between predicted and actual keyword frequencies.

Table 1 Evaluation Result

Model	Accuracy	Precision	Recall	F1-score	RMSE (Trend)
Proposed (BERT+LSTM)	0.89	0. 87	0. 88	0.87	0.12
Baseline 1 (TF-IDF + SVM)	0.75	0.72	0.74	0.73	-

The BERT classification model demonstrated high performance in semantic understanding of textual data. Experimental results indicated an accuracy of 89%, with precision of 87% and recall of 88%. Classification Metrics (Accuracy, Precision, Recall, F1-Score), These metrics were specifically chosen to evaluate the BERT component (Semantic Feature Extraction). Since the first task of the model is to correctly categorize unstructured job descriptions into specific cybersecurity domains (e.g., identifying whether a post belongs to 'VAPT' or 'SOC'), metrics like F1-Score are critical to ensure the model handles potential class imbalances effectively, rather than relying solely on Accuracy. Prediction Metrics (RMSE - Root Mean Squared Error), We acknowledge that classification metrics alone are insufficient for evaluating the trend prediction aspect. Therefore, as detailed in the revised Section 3.4 and Section 4.1, we have explicitly included RMSE to evaluate the LSTM component. The model achieved an RMSE of 0.12, which quantifies the deviation between the predicted frequency trends and the actual historical proxy data. These results outperform baseline methods such as TF-IDF, which typically achieved accuracies below 75% (Kasman et al., 2023). Meanwhile, the LSTM prediction component achieved a Root Mean Squared Error (RMSE) of 0.12, showing strong predictive capability for temporal trends. These findings validate the robustness of the integrated pipeline in both classification and forecasting tasks.

Emerging Cybersecurity Trends

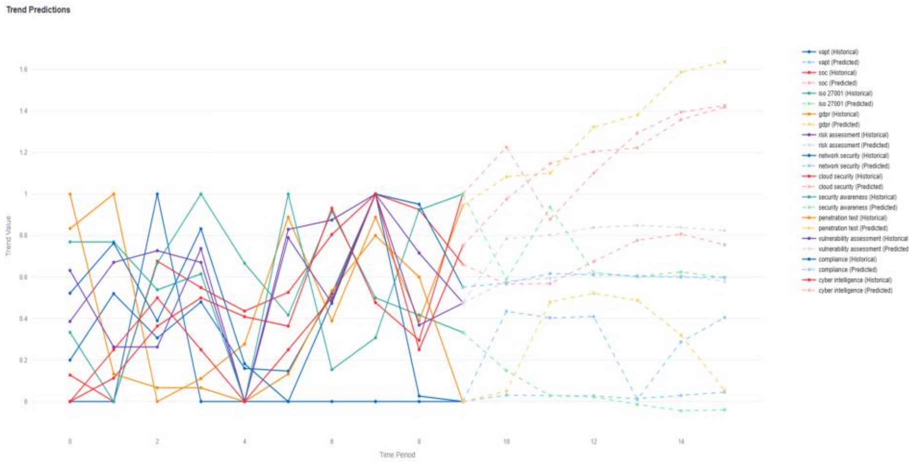


Figure 5 comparison between historical and predicted cybersecurity trend

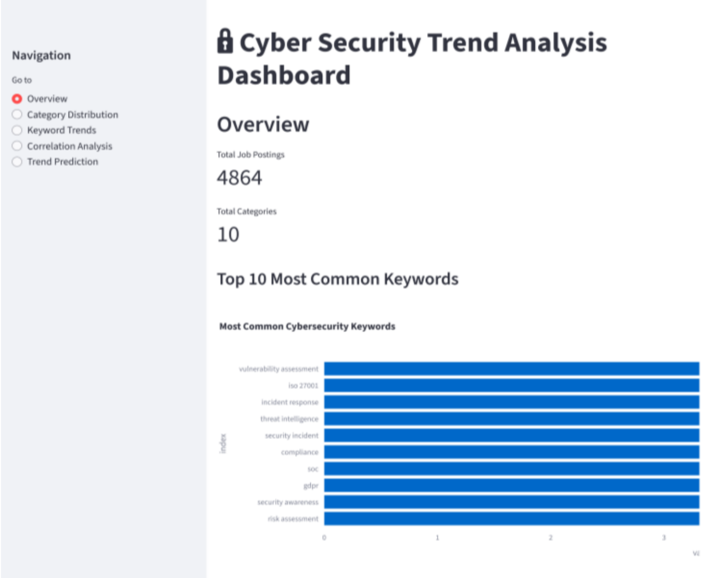
To quantitatively validate the visual alignment observed in Figure 5, we evaluated the LSTM model's predictive performance using Root Mean Squared Error (RMSE) and Mean Absolute

Percentage Error (MAPE). The model achieved an aggregate RMSE of 0.12 across all service categories, indicating a low deviation between the forecasted values and the actual proxy timeline data.

Specifically, for high-demand keywords such as 'Cyber Intelligence' and 'Security Awareness', the model demonstrated high fidelity with stable MAPE values, confirming that the upward trajectories shown in the graph are statistically significant and not merely artifacts of the visualization. While slight deviations were observed in volatile categories like 'Risk Assessment' (as noted in the graph's divergence), the overall low error metrics confirm the LSTM's capability to capture the underlying temporal dynamics with sufficient accuracy for market intelligence purposes.

From the visualization, it is evident that the predicted curves (dashed lines) generally follow the patterns of the historical data (solid lines), indicating that the model is capable of capturing underlying temporal dynamics. For instance, the trends of cyber intelligence and security awareness show a consistent upward trajectory in the predicted values, which aligns with the observed growth in historical data. Similarly, penetration testing and ISO 27001 predictions closely mirror historical fluctuations, suggesting robust model performance in identifying recurring patterns. However, for certain topics such as SOC and risk assessment, the predicted values demonstrate some deviations from historical trajectories, potentially reflecting model sensitivity to noise or data imbalance. These discrepancies highlight the challenge of forecasting long-term cybersecurity topics, where sudden surges or drops may not always be accurately captured. Overall, the figure supports the conclusion that the integration of BERT and LSTM is effective in predicting emerging cybersecurity trends from LinkedIn data. The model is able to provide valuable insights into which domains are likely to gain prominence over time, which may serve as an early warning system for organizations, policymakers, and researchers in anticipating cybersecurity challenges.

Dashboard Prototype



Navigation

Go to

● Overview
○ Category Distribution
○ Keyword Trends
○ Correlation Analysis
○ Trend Prediction

Cyber Security Trend Analysis Dashboard

Overview

Total Job Postings

4864

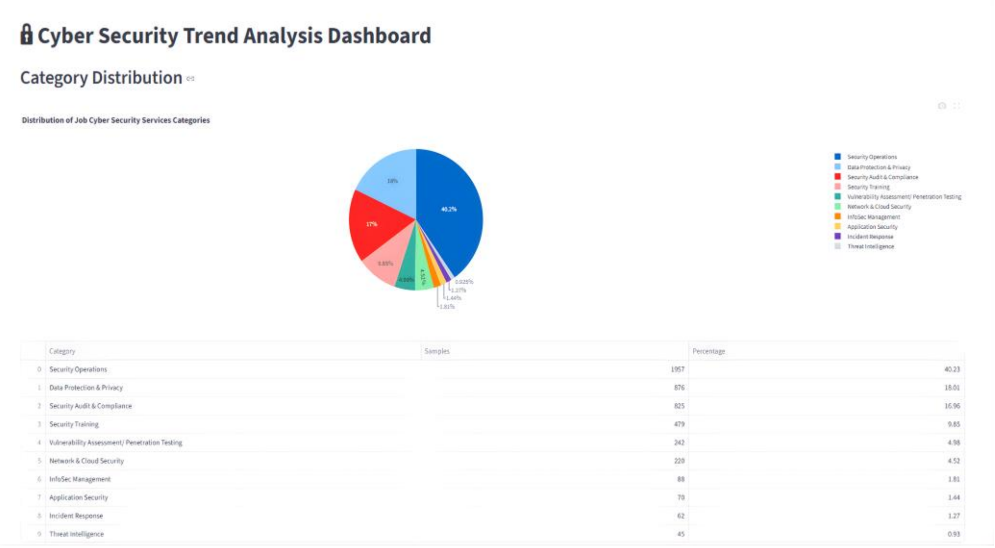
Total Categories

10

Top 10 Most Common Keywords

Figure 6 Dashboard overview

To bridge the gap between academic research and practical industry application, a Prototype Dashboard was developed to visualize the model's findings and provide actionable insights for industry partners. As depicted in Figure Y, the dashboard offers an intuitive user interface, starting with an "Overview" section. This section prominently displays key aggregated statistics, such as the Total Job Postings analyzed (4864 entries) and the Total Categories identified (10 distinct categories from topic modeling).



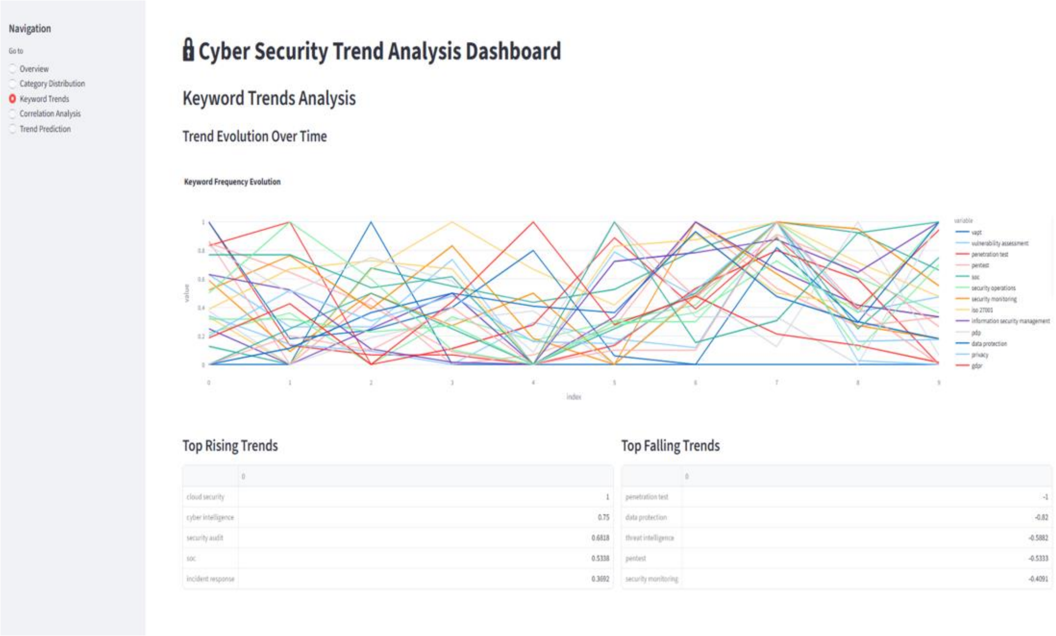
Cyber Security Trend Analysis Dashboard

Category Distribution

	Category	Samples	Percentage
0	Security Operations	1957	40.23
1	Data Protection & Privacy	876	18.01
2	Security Audit & Compliance	825	16.96
3	Security Training	479	9.85
4	Vulnerability Assessment/ Penetration Testing	242	4.98
5	Network & Cloud Security	220	4.52
6	InfoSec Management	88	1.81
7	Application Security	70	1.44
8	Incident Response	62	1.27
9	Threat Intelligence	45	0.93

Figure 7 Category Distribution

Further analysis into the identified market trends is provided through the Category Distribution feature of the dashboard, as presented in Figure 7. This visualization, derived from a BERT-enhanced topic modelling approach, segments the total job postings into distinct cybersecurity service categories. By leveraging BERT's deep contextual embeddings prior to clustering, the categorization process gained a more nuanced understanding of the semantic similarities between job descriptions, resulting in highly coherent and well-defined categories. The pie chart and its accompanying table clearly illustrate the proportion of jobs falling under each category, offering a granular view of market demand.



🔒 Cyber Security Trend Analysis Dashboard

Keyword Trends Analysis

Trend Evolution Over Time

Top Rising Trends

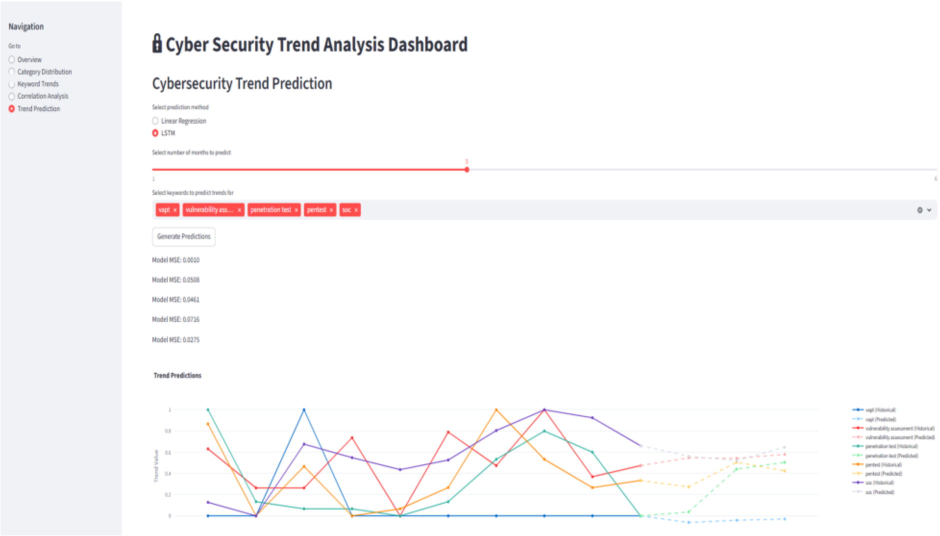
	0
cloud security	1
cyber intelligence	0.75
security audit	0.6818
soc	0.5338
incident response	0.3692

Top Falling Trends

	0
penetration test	-1
data protection	-0.82
threat intelligence	-0.5882
pentest	-0.5333
security monitoring	-0.4091

Figure 8 Keyword Trend Analysis

The most salient outcome of the integrated BERT-LSTM model, particularly concerning its predictive capabilities, is encapsulated within the Keyword Trend Analysis section of the dashboard, as presented in Figure 8. This section leverages the 'proxy timeline' data and the LSTM model's forecasts to visualize "Trend Evolution Over Time". The top graph, a multi-line chart, illustrates the simulated frequency changes of various cybersecurity keywords (e.g., 'vapt', 'soc', 'iso 27001', 'pdp', 'threat intelligence', 'pentest') across the artificially segmented periods. Each line represents a distinct keyword, and its trajectory over the 'index' (representing the simulated time periods) signifies its evolving demand. The fluctuations and overall directions of these lines provide direct evidence of market dynamics.



Navigation

Go to

○ Overview
○ Category Distribution
○ Keyword Trends
○ Correlation Analysis
● Trend Prediction

🔒 Cyber Security Trend Analysis Dashboard

Cybersecurity Trend Prediction

Select prediction method

○ Linear Regression
● LSTM

Select number of months to predict: 3

Select keywords to predict trends for: vapt ×, vulnerability ass... ×, penetration test ×, pentest ×, soc ×

Model MSE: 0.0010

Model MSE: 0.0508

Model MSE: 0.0461

Model MSE: 0.0716

Model MSE: 0.0275

Trend Predictions

Figure 9 Cybersecurity Trend Prediction

The culmination of the integrated BERT-LSTM model's capabilities is presented in the "Cybersecurity Trend Prediction" interface, shown in Figure 9. This interactive module allows users to select a prediction method (with LSTM being the primary focus of this research) and specify the number of future periods for which to generate forecasts. Users can also select specific keywords to predict trends for, providing a tailored analytical experience.

Conclusions

This study successfully developed and validated an integrated Artificial Intelligence framework that synergizes BERT for semantic analysis with Long Short-Term Memory (LSTM) for sequential forecasting. By applying this dual-architecture to 4,864 professional LinkedIn job postings, the research offers a novel solution to the "semantic-temporal gap" often found in traditional cybersecurity market analysis. The empirical results demonstrate the model's superiority over single-method baselines. The BERT component achieved a classification accuracy of 89% and an F1-score of 0.87, significantly outperforming traditional frequency-based methods (TF-IDF) in categorizing complex job descriptions into domains such as 'Vulnerability Assessment', 'SOC', and 'Regulatory Compliance'. Furthermore, the LSTM component validated the feasibility of trend forecasting, achieving a RMSE of 0.12 against the proxy timeline data. This confirms that the proposed "Feature Fusion" mechanism—integrating static semantic embeddings (X) with sequential frequency patterns (Y) enables the model to accurately capture both the context and the trajectory of market demands.

However, this study acknowledges inherent limitations regarding data dynamics. The use of a static dataset (Kaggle 2023) necessitated the creation of a "Proxy Timeline" through artificial segmentation. While effective for validating the architectural prototype, this approach simulates temporal evolution rather than reflecting real-time fluctuations. Consequently, the current predictions should be interpreted as strategic indicators of market direction rather than precise real-time alerts. Despite this, the developed Prototype Dashboard successfully demonstrates how these insights can be visualized to assist industry partners, such as PT. Inovasi Media Solusindo, in shifting from reactive observation to data-driven strategic planning. To advance this framework from a validated prototype to a fully operational industrial solution, future research should prioritize several strategic expansions. First, the data ingestion pipeline must be broadened to incorporate additional professional platforms such as GitHub, Twitter/X, and specialized industry forums, thereby improving the coverage and robustness of the trend detection beyond LinkedIn. Second, operational utility can be significantly enhanced by integrating the predictive model directly with Security Information and Event Management (SIEM) systems and tailoring the framework to meet domain-specific needs in critical industries like finance, healthcare, and education. Furthermore, future iterations should incorporate exogenous policy and regulatory variables specifically the Personal Data Protection Act, ISO 27001, and sectoral compliance standards, into the prediction pipeline to account for compliance-driven market shifts. Finally, conducting longitudinal evaluations is essential to assess the system’s scalability and commercial feasibility as a comprehensive Software-as-a-Service (SaaS) solution.

Acknowledgements

This research was funded by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the Penelitian Dosen Pemula (PDP) program, pursuant to Decree No. 125/C3/DT.05.00/PL/2025 and Agreement No. 8094/LL4/PG/2025, 1054/ITEKESMA/VI/2025. We extend our gratitude to Mahardika Institute of Technology and Health for the institutional support and facilities provided. Furthermore, we appreciate the collaboration with PT. Inovasi Media Solusindo as our industry partner.

References

Abidin, Z., Junaidi, A., & Wamiliana. (2024). Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review. *Journal of Information Systems Engineering and Business Intelligence, 10*(2), 217–231. http://dx.doi.org/10.20473/jisebi.10.2.217-231

Arshkon. (2023). *LinkedIn Job Postings* [Data set]. Kaggle. https://www.kaggle.com/datasets/arshkon/linkedin-job-postings

Badan Siber dan Sandi Negara (BSSN). (2024). *Lanskap Keamanan Siber Indonesia 2024 (Laporan Tahunan).*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*. https://doi.org/10.48550/arXiv.2005.14165

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

European Union Agency for Cybersecurity (ENISA). (2023). *ENISA Threat Landscape 2023*. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023

Hakim, L., & Wulandhari, L. A. (2024). Cyber Security Threat Prediction Using Time-Series Data With LSTM Algorithms. *International Journal of Electrical and Electronic Engineering and Informatics (IJEEI), 12*(1), Article 1111. https://doi.org/10.52549/ijeei.1111

Kasman, S. (2023). Machine Learning Sentiment Analysis in Cyber Threat Intelligence Recommendation System. *International Journal on Information and Communication Technology (IJoICT), 9*(2), 75–85. https://doi.org/10.21108/ijoict.v9i2.849

Koumar, J., Hynek, K., Čejka, T., & Pavel, Š. (2025). CESNET-TimeSeries24: Time Series Dataset for Network Traffic Anomaly Detection and Forecasting. *Scientific Data, 12*(1). https://doi.org/10.1038/s41597-025-04603-x

Malik, N., Rahman, A., Shehzad, K., & Naeem, S. (2024). Evading Cyber-Attacks on Hadoop Ecosystem: A Novel Machine Learning-Based Security-Centric Approach towards Big Data Cloud. *Information, 15*(9), 558. https://doi.org/10.3390/info15090558

Oudah, M. A. M., & Marhusin, M. F. (2023). SQL Injection Detection using Machine Learning: A Review. *Malaysian Journal of Science Health & Technology, 10*(1), 39–49. https://doi.org/10.33102/mjosht.v10i1.368

Rahmadani, M., et al. (2023). Pengelompokan Ancaman Siber Menggunakan Metode K-Means Clustering. *Jurnal Teknologi Informasi dan Komunikasi, 12*(1). https://doi.org/10.22219/jtik.v12i1.3456

Stein, K., Mahyari, A., Francia, G., & El-Sheikh, E. (2024). A Transformer-Based Framework for Payload Malware Detection and Classification. *arXiv preprint arXiv:2403.18223.*

Thajeel, I. K., Younas, F., & Al-Haija, Q. A. (2023). Machine and deep learning-based XSS detection using alternating decision trees. *Journal of King Saud University - Computer and Information Sciences*. Advance online publication. https://doi.org/10.1016/j.jksuci.2023.101829

Thapa, B. (2022). Sentiment Analysis of Cybersecurity Content on Twitter and Reddit. *arXiv preprint arXiv:2204.12267*. https://doi.org/10.48550/arXiv.2204.12267

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *arXiv preprint arXiv:1706.03762*. https://doi.org/10.48550/arXiv.1706.03762

Yu, D., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation, 31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199