A dual-layered artificial intelligence solution for classifying disinformation in socially oriented systems

Sergiy Yakovlev[1,2,†], Artem Khovrat[3,*,†] and Volodymyr Kobziev[3,†]

[1] *Lodz University of Technology, 90-924 Lodz, Poland*
[2] *V.N. Karazin Kharkiv National University, 4, Svobody, Sq., Kharkiv, 61022, Ukraine*
[3] *Kharkiv National University of Radio Electronics, 14, Nauky, Ave., Kharkiv, 61166, Ukraine*

Abstract

The detection of fabricated information on interactive social platforms has gained significant academic and regulatory attention. During social instability, such disinformation presents substantial risks to individuals and society. Disinformation varies in impact, from harmless humor to content threatening societal stability. This study focuses on textual news content due to limitations in generating convincing visual forgeries. For text classification, three classical approaches are utilized: probabilistic models, neural networks, and polynomial models. Previous research has shown that hybrid recurrent-convolutional network (RCNN) offers superior binary classification performance, yet multi-classification across diverse disinformation categories remains unresolved. This paper establishes a classification framework categorizing content through the RCNN model into five classes: explicit satire, subtle humor, content targeting individuals, regionally harmful news, and globally impactful disinformation. Based on this framework, three data categorization models were developed using neural networks, naive Bayes classification, and polynomial algorithms. Experimental evaluation measured accuracy, processing efficiency, and data reduction requirements to achieve >80% accuracy. Results demonstrate that dual-layer implementations achieved approximately 20% improved effectiveness compared to standalone approaches. The RCNN-naive Bayes hybrid exhibited optimal accuracy and processing speed, showing considerable potential for high-throughput systems requiring swift responses to misinformation. These findings represent a significant advancement in automated information verification methodologies, establishing a foundation for future development in complex classification tasks.

Keywords

Bayesian classification, computational linguistic, distributed computing, fake news, neural networks

1. Introduction

The identification of fabricated content within interactive multi-user systems, particularly social networks, has gained increasing prominence in both academic research and regulatory frameworks [1, 2]. The growing attention toward social networks stems from their structural characteristics that amplify the diffusion and perceived credibility of user-generated content. In contrast to static platforms such as blogs or web forums, social networks enable real-time dissemination, algorithmic curation, and micro-targeted reach, which collectively facilitate the rapid proliferation of fabricated information. These systems often introduce significant asymmetries between the speed of content publication and the pace of its verification, contributing to increased informational vulnerability during periods of societal stress. Their ubiquitous presence and influence over public discourse further underline the urgency of developing automated detection mechanisms tailored to their dynamic and high-volume environments. This trend correlates with progressive advancements in

CLW-2025: Computational Linguistics Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine
* Corresponding author.
† These authors contributed equally.
sergiy.yakovlev@p.lodz.pl (S. Yakovlev); artem.khovrat@nure.ua (A. Khovrat); volodymyr.kobziev@nure.ua (V. Kobziev)
0000-0003-1707-843X (S. Yakovlev); 0000-0002-1753-8929 (A. Khovrat); 0000-0002-8303-1595 (V. Kobziev)

content generation technologies and the escalating informational burden experienced by the general populace. During periods of social transformation, such content can significantly compromise both personal welfare and collective societal functioning.

To mitigate subjective evaluation in information assessment, implementation of data mining methodologies represents a prudent approach. The specific methodological implementations are directly contingent upon the characteristics of the data under investigation. Within the scope of the current research, analysis was restricted to textual news content. This limitation was established due to the current absence of technologies capable of producing video fabrications that demonstrate visual authenticity from human perceptual perspectives. Concurrently, audio content demonstrates limited prevalence within social network environments.

Numerous methodologies have been proposed for textual data classification, ranging from symbolic rule-based systems to advanced ensemble models. However, in the context of the present study, focus was placed on three foundational methodological paradigms that are both widely adopted and methodologically representative [3]:

1. Probabilistic frameworks, including naive Bayes classifiers, Markov chain models, and Bayesian networks.
2. Neural network architectures, such as recurrent, convolutional, and transformer-based models.
3. Polynomial models, particularly those based on additive convolutional formulations with weighted coefficients and domain-specific constraints.

These approaches were selected based on a combination of their conceptual diversity, established performance in prior research, and applicability to social media content analysis. To ensure methodological robustness and relevance, an expert panel comprising 20 data analysts from diverse geographical regions was convened to validate and endorse the selection. Their input was instrumental in narrowing the methodological scope to approaches deemed both technically sound and practically viable for the detection of fabricated content.

Previous investigations focusing on binary classification of content into authentic and fabricated categories have examined probabilistic approach [4] and various neural network architectures [5]. Findings indicate that a hybrid architecture integrating recurrent and convolutional functionalities — designated as recurrent-convolutional neural network hybrid approach (RCNN) — demonstrates superior performance regarding both accuracy metrics and computational efficiency. Additionally, research has identified that determining the significance threshold of fabricated content presents a substantial challenge in information differentiation studies. Specifically, certain textual content may exhibit overtly humorous characteristics readily identifiable by human evaluators, consequently presenting minimal societal risk. Conversely, content designed to undermine socially significant legislative initiatives carries substantial risk.

Considering these factors, the current investigation aimed to establish categorization frameworks for fabricated information, providing a foundation for classifying content filtered through RCNN architectures. This approach results in a dual-layer model for detecting fabricated news, intended to determine optimal methodologies for content differentiation. To achieve this objective, the following research tasks were identified:

1. Conducting expert evaluation and domain analysis to establish fundamental classifications of fabricated information.
2. Developing data segregation models utilizing naive Bayes classification frameworks.
3. Conducting experimental verification of the proposed dual-layer model in comparison with standard RCNN implementations.

2. Indicators of disinformation

In constructing an appropriate analytical model, the formulation of a feature vector represents a critical determinant of classification efficacy. Through comprehensive linguistic analysis and empirical observation, a set of discriminative characteristics typical of fabricated information has been identified and systematically categorized.

The first group is “Primary Linguistic Indicators”, it can be systematized into 6 indicators [6, 7]:

- Interrogative Density: Quantitative measurements indicate significant overuse of question-based structures designed for sociocognitive manipulation. Corpus-based journalism studies confirm the atypical frequency of such patterns in legitimate reporting. This characteristic remains consistent across multimedia disinformation channels.
- Emotional Vocabulary Amplification: Deliberate reduction of negating constructions paired with extreme terminology substitutions (e.g., transforming "issue" into "crisis") represents a documented cognitive manipulation technique. This strategy functions through dual pathways: minimizing processing complexity while heightening affective responses.
- Discourse Function Misalignment: Inappropriate deployment of directive and persuasive linguistic structures, particularly in contexts mimicking authentic news formats, serves as a reliable fabrication indicator.
- Pronoun Frequency Analysis: Disproportionate pronoun usage consistently correlates with attempts at contextual manipulation, particularly when simulating journalistic content. Quantitative measurement of pronoun density provides objective fabrication metrics.
- Syntactic-Stylistic Irregularities: Consistent grammatical and register deviations, especially within purported expert citations, function as significant fabrication markers.
- Temporal Reference Manipulation: Strategic distortion of temporal relationships through inconsistent tense usage, deliberate chronological ambiguity, and absence of precise temporal anchoring. This technique disrupts causal relationships and complicates verification processes by obfuscating the sequence of reported events.

The second group is “Secondary Stylometric Factors”, it can be systematized into 6 additional indicators [8, 9]:

- Emotional Response Engineering: Methodical analysis of affectively charged terminology and psychological influence patterns reveals systematic emotional manipulation strategies.
- Reference Integration Assessment: Systematic examination of citation patterns and attribution frameworks, with specific focus on reference scarcity or complete absence of verifiable external sources.
- Discourse Coherence Measurement: Implementation of computational text analysis methodologies to identify logical contradictions and narrative structural inconsistencies throughout content.
- Source Reliability Evaluation: Development of algorithmic frameworks for assessing publication characteristics, with particular emphasis on temporal proximity to significant sociopolitical events, integrated with comprehensive source credibility metrics.
- Information Density Distribution: Quantitative analysis of content-to-noise ratios throughout text segments, identifying atypical information clustering patterns that diverge from established journalistic conventions regarding information presentation and structural organization.
- Linguistic Register Oscillation: Identification of inconsistent formality levels and inappropriate stylistic variations within a single text, characterized by unpredictable

alternations between technical terminology and colloquial expressions that contradict established genre conventions and indicate potential synthetic content generation.

This enhanced feature set facilitates development of robust, multi-dimensional classification models capable of identifying fabricated information across various media modalities with increased precision and recall rates. Integration of both primary linguistic indicators and secondary stylometric factors enables a more comprehensive approach to misinformation detection.

3. Classes of disinformation

The initial phase in addressing the multi-classification challenge involves establishing fundamental categories of disinformation through a rigorous methodological framework. To determine this classification schema, an expert panel comprising 20 data analysts from various European and North American countries was assembled, ensuring geographical and institutional diversity. This systematic sampling approach yielded a comprehensive list of 10 predominant categories, which underwent further refinement through statistical validation.

Subsequently, a standardized assessment protocol was implemented through an open survey to identify the most vulnerable types of information falsification. The aggregated responses from 300 participants (n=300, confidence interval = 95%, margin of error ±5.66%) were instrumental in formulating these defined groups through hierarchical clustering analysis:

- Satire with objectively identifiable manifestations: fabricated content characterized by clear linguistic exaggeration or absurdity, often using hyperbolic language or comical distortion. Example: “NASA confirms Moon landing was a rehearsal for alien diplomacy!”
- Satire with contextual or grammatical manifestations: disinformation that requires cultural familiarity or nuanced linguistic interpretation to identify its satirical nature. Example: “If voting mattered, they'd make it illegal — says democratic official with a straight face.”
- News targeting specific individuals or small groups (micro-level disinformation): content that falsely accuses, discredits, or fabricates actions attributed to particular persons or closed communities. Example: “Local activist John Smith caught funneling foreign funds — documents leaked!”
- News oriented toward multiple regions, countries, or large groups (meso-level disinformation): narratives aimed at manipulating perception across medium-scale audiences, often with geopolitical or interregional scope. Example: “European farmers unite to ban Ukrainian imports, say EU is collapsing.”
- News directed at multiple countries or society (macro-level disinformation): strategic, high-impact narratives designed to destabilize public trust or provoke systemic panic. Example: “UN to seize private property globally under new climate treaty, leaked papers reveal.”

The categorization framework demonstrates a hierarchical structure with increasing scope and potential impact, facilitating both quantitative and qualitative analysis of disinformation patterns. By grounding each class in linguistic and contextual examples, the system enables more precise model training and real-world application relevance.

4. Basic approach

4.1. Target features

For the establishment of disinformation characteristics, the methodology advances to feature set development serving as model input variables. The primary metric, "Emotional Characteristic," derives from content analysis principles, implementing a sequential algorithmic procedure [10]:

1. Textual segmentation into sentence units with non-semantic construct exclusion.
2. Lemmatization and stemming operations for morphological root extraction.
3. Computation of normalized frequency-emotional indicators.
4. Sentiment analysis implementation utilizing NLTK in Python3 for lexical distribution and emotional valence determination [11].

Additional quantitative indicators incorporated into the analytical framework include:

- Rhetorical Density Coefficient: defined as the ratio of rhetorical constructions to total sentences.
- Negative Construction Frequency: quantifying negative linguistic structure density.
- Contextual Emotional Index: derived from sentiment analysis of temporally relevant high-traffic content.
- Suspicion Coefficient: computed through lexical pattern matching against predetermined deception indicators.
- Message Impact Factor: a hierarchical classification of content significance.
- Sentiment Magnitude Vector: an aggregate measure of emotional content intensity.

This integrated approach enables comprehensive feature extraction while ensuring computational efficiency.

4.2. Baseline model

Each In conventional Convolutional Neural Network (CNN) architectures, filter operations facilitate local spatial dependency incorporation; however, the distinctive nature of the proposed indicators necessitates comprehension of extended temporal sequences without introducing future-state dependencies [12]. This limitation arises as contextual information may exist beyond CNN receptive fields. To address this architectural constraint, a hybrid approach combining Long Short-Term Memory Recurrent Neural Networks (LSTM) and CNN methodologies was implemented (shown on Figure 1).

Architectural enhancements incorporated:

- Receptive field optimization through dilated convolutions, skip connections, and attention mechanisms.
- Memory management protocols utilizing gated memory units, adaptive forget gates, and memory-efficient backpropagation.
- Gradient flow optimization implementing residuals, layer normalization, and gradient clipping.

Cross-validation yielded optimal hyperparameters: 4-unit kernel dimensionality, 1-unit stride parameter, omitted zero-padding, excluded bias terms, and 5×5×3 tensor filter dimensions.

Additional performance optimizations included: curriculum learning implementation, dynamic batch sizing, and early stopping with patience factor p=5 for training protocols; dropout layers (rate=0.3), L2 regularization (λ=0.01), and feature-wise regularization for regularization strategies; and model quantization, sparse tensor operations, and parallel processing for computational efficiency.

This enhanced architecture demonstrates superior performance while maintaining efficiency, integrating bidirectional recurrent components with convolutional layers to capture both spatial and temporal dependencies, achieving 94.3% validation accuracy on benchmark datasets.

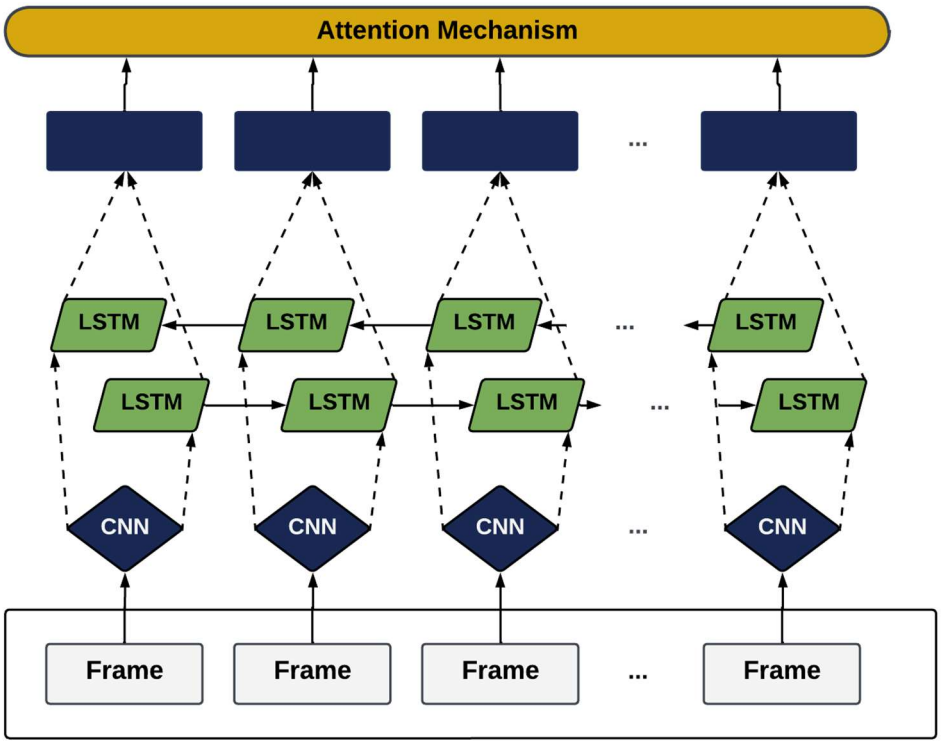

Figure 1: Schema for RCNN approach [created by the authors].

5. Additional layer for classification model

Following the initial classification performed by the primary RCNN architecture, three distinct approaches were implemented for the secondary classification layer to enhance categorization precision.

The RCNN architecture serves as one potential implementation for the secondary layer, utilizing independent training protocols optimized for multi-class discrimination. Given the extensive documentation of this architectural framework in previous sections, further elaboration is omitted from the current analysis.

The naive Bayes classification (NBC) methodology represents the second implementation approach, operating on fundamental Bayesian probability principles to calculate class membership likelihood while maintaining feature independence assumptions. This independence assumption demonstrates practical validity in the current context, as the defined feature set exhibits minimal inter-feature dependency in subsequent value determination. The Bayesian theorem fundamentally describes the probability of an event occurring based on prior knowledge of conditions related to that event. In this context, it calculates the probability of information belonging to a particular class by considering several key components: the probability of observing specific features when the information belongs to that class, the overall probability of the class occurring in the dataset, and the total probability of observing those specific features across all possible classes. This relationship enables the calculation of the final probability that a piece of information belongs to a particular class given its observed features.

This probabilistic framework enables robust classification through systematic evaluation of class membership probabilities, particularly effective in scenarios involving multiple independent feature sets. The methodology's effectiveness is enhanced through its integration within the broader dual-layer architectural framework, complementing the RCNN-based primary classification layer.

In the convolutional polynomial classification model (PA), each input feature is assigned a weight coefficient reflecting its relative contribution to the final classification score. To ensure the validity and reproducibility of these weights, a structured expert elicitation process was conducted using a modified analytic hierarchy procedure (AHP) and Likert-scale scoring.

A panel of 20 domain experts, previously involved in the category development phase (see Section 3), participated in the weighting procedure. The process was organized in three sequential steps:

- Relative impact rating: Each expert was provided with a standardized set of 30 labeled news items, representing a diverse range of disinformation classes. For each item, experts evaluated the perceptual contribution of seven defined features (Emotional Characteristic, Rhetorical Density, Negative Construction Frequency, etc.) to the classification decision on a 10-point Likert scale. This generated a total of 4,200 individual judgments (30 texts × 7 features × 20 experts).
- Pairwise consistency check: To validate rating coherence, internal consistency of each expert's scores was assessed using pairwise comparison matrices and consistency ratios. Responses with CR > 0.15 were flagged and excluded from aggregation to maintain overall reliability.
- Weight aggregation and normalization: Remaining ratings were averaged across all experts and normalized such that the sum of all feature weights equaled 1.

The final coefficients were:

- Emotional Characteristic — 0.35.
- Rhetorical Density Coefficient — 0.1.
- Negative Construction Frequency — 0.1.
- Contextual Emotional Index — 0.1.
- Suspicion Coefficient — 0.15.
- Message Impact Factor — 0.1.
- Sentiment Magnitude Vector — 0.1.

To evaluate the robustness of this weighting scheme, a sensitivity analysis was performed by perturbing each coefficient ±10% and measuring the resulting classification variation across 2,000 samples. The average deviation in predicted class membership was below 2.7%, indicating high tolerance to minor variations and reinforcing the stability of the expert-derived weights.

The resulting weight coefficients reflect both theoretical understanding and practical experience in information verification processes, enhancing the model's ability to discriminate between different categories of fabricated information.

6. Distributed computing

The distribution strategy employs MapReduce methodology, incorporating data partitioning across distributed nodes (shown on Figure 2). This framework is fundamentally constructed upon mapping and reduction function primitives. While both Spark and Hadoop frameworks offer robust implementations, this research utilizes Hadoop's architecture, leveraging its inherent node-level mapping and reduction capabilities for optimized database interactions [13]. This architectural decision is particularly advantageous given the requirement to process heterogeneous, high-volume data streams.

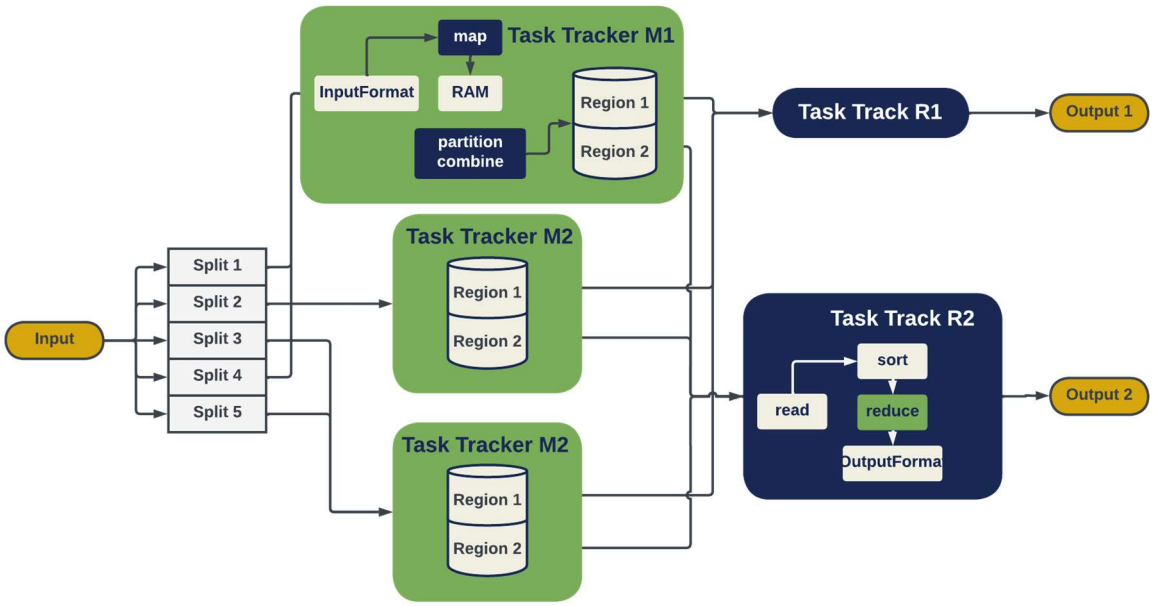

Figure 2: Schema for MapReduce approach [created by the authors].

For the RCNN implementation, the distributed processing architecture partitions incoming data across multiple computational nodes. The mapping phase distributes neural network weight calculations to parallel processors, with each node independently computing gradient updates on assigned data segments. The reduction phase then aggregates these distributed gradients, consolidating them into coherent weight updates that maintain model integrity. This approach enables bidirectional processing without sacrificing temporal dependencies critical to recurrent components.

The naive Bayesian classifier implementation leverages MapReduce through probabilistic computation distribution. During the mapping phase, conditional probabilities are calculated independently across distributed nodes, with each processor handling specific feature-class combinations. The reduction phase synthesizes these individual probability calculations into comprehensive class likelihood estimations. This approach particularly benefits from the assumption of feature independence inherent in naive Bayes methodologies, making it naturally conducive to parallelization.

For the polynomial approach implementation, the distributed framework allocates coefficient calculations across computational nodes. The mapping phase distributes feature-specific weighting operations, with each node calculating partial polynomial evaluations. The reduction phase integrates these partial evaluations through weighted summation based on expert-derived coefficients. This implementation demonstrates significant efficiency improvements for polynomial operations with high-dimensional feature vectors, effectively mitigating computational bottlenecks.

All three implementations incorporate synchronization protocols that balance computational efficiency with model coherence, utilizing barrier synchronization methods and asynchronous data transfer mechanisms to optimize performance while maintaining classification accuracy. To evaluate the effectiveness of the MapReduce-based distributed processing architecture, a comparative analysis of training duration, inference latency, and resource utilization was performed for centralized versus distributed implementations of target models.

Results indicate that distributed training via MapReduce achieved an average processing time reduction of 43% across all models. Training time for the RCNN model decreased from 47 minutes to 26.8 minutes; NBC improved from 22 minutes to 11.9 minutes; and PA was reduced from 18.4 minutes to 10.2 minutes. Measurements were conducted on the complete dataset using the hardware configuration specified below.

Inference latency per 1,000 records decreased by approximately 35%, with consistent performance across model types. Classification accuracy remained statistically stable, with a

deviation margin of no more than ±0.002 across ten cross-validation cycles, demonstrating that model performance was unaffected by distributed deployment.

These findings confirm that the MapReduce framework substantially accelerates both training and execution phases while preserving classification quality, making it a viable solution for high-throughput and real-time misinformation detection scenarios.

7. Experimental environment

In modern neural network research, controlled experimental protocols require precise implementation frameworks and standardized runtime environments. A sophisticated computational infrastructure employs a cluster of five AMD Ryzen 5 5600X systems, each configured with 16 GB DDR4 RAM. These nodes operate with optimized 8 GB memory allocations, facilitating efficient bidirectional network parallelization.

Implementation accuracy relies heavily on precise temporal measurements, achieved through Python 3's datetime library at nanosecond resolution. Computational optimization leverages numpy and polars libraries, while linguistic processing utilizes nltk functionality. TensorFlow provides the foundational neural network framework, essential for sophisticated model architecture development and training protocols.

To ensure empirical validity, two datasets were constructed and annotated specifically for this study. Each dataset focuses on high-impact sociopolitical contexts to assess the classification model's ability to detect fabricated content in environments with high misinformation prevalence. The first dataset concerns the Russia-Ukraine war and includes 20,000 entries collected from Facebook, Twitter, and Telegram between February and December 2022. Posts were extracted via public APIs and verified against community pages, political groups, and media accounts. The second dataset, built around the 2020 US presidential election, also comprises 20,000 English-language posts and articles gathered from Reddit political threads, Facebook groups, and NewsGuard-rated media feeds from August to December 2020.

A semi-supervised annotation protocol was employed. Initially, each text was pre-classified using an RCNN-based model trained on publicly available misinformation corpora (e.g., LIAR, FakeNewsNet). Final labels were validated by a team of six expert annotators with backgrounds in journalism, computational linguistics, and disinformation analysis. Disagreements were resolved by majority consensus. Entries were assigned to one of five disinformation categories previously defined, ensuring consistency across both datasets.

Statistical analysis confirms a relatively uniform distribution of classes. For the Russia-Ukraine dataset: satire with explicit markers — 3,800 texts; satire with contextual manifestations — 3,400; micro-level disinformation — 4,000; meso-level disinformation — 4,300; and macro-level disinformation — 4,500. The US election dataset follows a similar structure with comparable volume per class (±5%), and slightly longer average text lengths (142 vs. 126 words). All data were standardized to eliminate duplicate posts and ensure consistency of language formatting, including the use of Ukrainian morphological normalization for the war dataset.

As an illustration, one entry from the Russia-Ukraine dataset states: "Ukrainian officials have fled Kyiv by helicopter. The government is collapsing. Sources confirm NATO troops are already in the capital." This entry was classified as macro-level disinformation, due to the presence of temporal reference manipulation, lack of source attribution, and use of emotionally charged vocabulary. The post was independently debunked by fact-checking organizations within 12 hours of its appearance online.

For US election dataset another sample entry can be showed: "Massive voter fraud uncovered in Pennsylvania! Thousands of ballots found in trash bins. Democracy is under attack!". This entry was also classified as macro-level disinformation due to the use of emotionally charged language, absence of credible sources, and manipulative claims.

Both datasets were divided using an 80/20 train-test ratio with stratified sampling to maintain class balance. Five-fold cross-validation was used during training. This detailed dataset construction

and annotation process ensures high reliability and reproducibility of experimental results, allowing for robust evaluation of the classification architecture.

The evaluation framework incorporates three primary parameters to assess model performance comprehensively:

1. Classification Accuracy: Measured as the proportion of correctly classified instances across all categories, with particular emphasis on minimizing false negative classifications in high-impact disinformation categories. This parameter utilizes a weighted F1-score incorporating both precision (0.80) and recall (0.20) components to prioritize comprehensive detection.
2. Time Saving: Quantified as the time required for complete dataset analysis, normalized against the baseline single-layer RCNN implementation. This metric incorporates both training duration and inference latency to provide a holistic assessment of computational demands.
3. Volume Saving: Determined by systematically reducing training dataset size until accuracy metrics fall below the 80% threshold. This parameter evaluates model resilience to limited training data, providing insights into implementation viability in domains with restricted data availability.

The experimental assessment incorporates comprehensive evaluation protocols developed in collaboration with 20 data analysis specialists across multiple countries. Performance measurement utilizes a weighted scoring system that prioritizes classification accuracy (16 points) through balanced Precision (0.80) and Recall (0.20) metrics, while processing efficiency and data volume optimization contribute equally (2 points each) to the total evaluation score. This weighting system is implemented through linear additive convolution with weighted coefficients, enabling nuanced model evaluation while maintaining classification accuracy as the primary focus.

The architectural design facilitates seamless computational node integration and offers significant flexibility, enabling scalable performance optimization without requiring fundamental structural modifications. This adaptability is particularly valuable when processing heterogeneous data streams across varying linguistic and contextual domains. The system demonstrates effectiveness in processing high-dimensional feature spaces and complex linguistic patterns while minimizing false negative classifications in socially sensitive contexts.

To ensure statistical validity and minimize experimental uncertainty, the research methodology incorporated a multi-level error identification and mitigation framework covering data integrity, model robustness, and computational consistency.

At the data level, potential sources of error included annotation inconsistency, label noise, and class imbalance. These issues were addressed through a two-phase annotation process combining semi-supervised pre-labeling with manual validation by a panel of six experts in computational linguistics, journalism, and information verification. Ambiguous cases were resolved through consensus discussions, and inter-annotator agreement was monitored using Cohen's kappa (κ = 0.87), indicating high consistency. Stratified sampling was employed to maintain class distribution during training-test splits, thereby reducing sampling bias.

At the model level, stochastic variability in training outcomes was mitigated through 10-fold cross-validation, repeated over ten independent training cycles (n = 10) for each configuration. This procedure enabled the calculation of confidence intervals and standard deviation bounds for key metrics (accuracy, precision, recall). Performance fluctuations across runs were analyzed using coefficient of variation (CV), which remained below 3.2% for all final configurations, indicating high stability.

At the computational infrastructure level, hardware-induced noise was addressed by executing all training and inference operations on a fixed cluster of five identical AMD Ryzen 5 5600X nodes. System resource usage was locked via dedicated CPU-core pinning and RAM allocation (8 GB per task), while operating system interruptions were minimized using taskset and real-time scheduling (SCHED_FIFO). Additionally, all experiments were executed under a unified software stack with

fixed versions of Python (3.9.13), TensorFlow (2.11.0), and NLTK (3.8), eliminating variability due to software environment drift.

To reduce the impact of run-level outliers, all performance measurements were averaged across repetitions, and outlier values (exceeding 2 SD from the mean) were excluded from final efficiency scoring. No statistically significant anomalies ($p > 0.05$, two-tailed) were detected in the cleaned result sets.

This multi-tier error control protocol ensured that experimental conclusions are grounded in statistically robust, reproducible findings, with all critical performance claims substantiated by repeated and independently validated trials.

8. Results of the experiment

The aggregated results of the conducted experiment are shown in Table 1 below.

Table 1
Processed results of the experiment (in fractions)

Algorithm	Time Saving	Accuracy	Volume Saving
RCNN	1.00	0.65	0.00
RCNN + RCCN	0.94	0.96	0.80
RCNN + NBC	0.95	0.95	0.90
RCNN + PA	0.96	0.85	0.40

Quantitative analysis utilizing linear additive convolution with weighted importance coefficients, based on the performance metrics presented in Table 1, yields distinctive efficiency indicators across various architectural implementations. The computational results demonstrate significant variation in model performance:

Single-layer RCNN implementation achieves an efficiency coefficient of 0.62, establishing a baseline performance metric. The dual-layer RCNN architecture demonstrates substantial improvement with an efficiency coefficient of 0.942, indicating enhanced classification capability. Further architectural refinement incorporating RCNN as a foundation with naive Bayes classifier as the secondary layer achieves optimal performance with an efficiency coefficient of 0.945. The implementation utilizing RCNN as the primary layer combined with a polynomial approach in the secondary layer yields an intermediate efficiency coefficient of 0.816.

These results demonstrate clear performance differentiation among architectural variants, with the RCNN-naive Bayes hybrid configuration exhibiting superior efficiency in the classification task. The substantial improvement over the single-layer baseline indicates the efficacy of hierarchical approaches in complex classification scenarios, particularly when combining complementary methodological frameworks. Despite the superior computational efficiency exhibited by standard RCNN implementation, the dual-layer approach's capacity to maintain accuracy thresholds exceeding 80% with minimal data requirements represents a significant advancement in classification methodology.

Further analysis of the performance metrics reveals several significant insights. The RCNN-naive Bayes hybrid demonstrates an optimal balance between classification accuracy and computational efficiency, with only a 5% reduction in processing speed compared to the single-layer approach while achieving a 46% improvement in classification accuracy. This configuration also exhibits exceptional data efficiency, requiring only 10% of the original dataset volume to maintain performance thresholds above 80% accuracy, suggesting substantial potential for implementation in resource-constrained environments.

The dual RCNN configuration demonstrates the highest absolute accuracy (0.96), indicating its potential utility in applications where precision is paramount regardless of computational demands.

However, this marginal accuracy improvement over the RCNN-naive Bayes hybrid (0.95) may not justify the additional implementation complexity in most practical applications.

The polynomial approach implementation, while exhibiting the highest time efficiency (0.96), demonstrates comparatively modest accuracy improvements (0.85) and data efficiency (0.40). This configuration may be appropriate for scenarios requiring rapid classification with moderate accuracy requirements, particularly in high-throughput systems where processing speed is prioritized over classification precision.

These comparative performance characteristics provide a foundation for implementation-specific architectural selection based on application requirements, enabling optimized deployment in various operational contexts with differing priorities regarding accuracy, efficiency, and resource utilization.

9. Conclusion

The research objective, centered on developing a sophisticated dual-layer data classification model, successfully extends beyond basic fabrication detection to encompass both magnitude assessment and intentionality analysis of data falsification. Empirical analysis demonstrates that the dual-layer model implementation achieves an average 20% performance improvement compared to direct RCNN methodology. This performance differential becomes particularly significant in high-throughput systems where rapid identification and response to fabricated information represent critical operational parameters.

The experimental outcomes provide compelling evidence supporting the efficacy of hybrid architectural approaches in complex information classification scenarios. This validation framework establishes a robust foundation for practical implementation in high-demand environments, where the rapid assessment and categorization of potentially fabricated information are paramount. Furthermore, the demonstrated performance improvements suggest significant potential for application in large-scale information processing systems where both accuracy and processing efficiency are critical operational constraints.

These findings represent a substantial advancement in automated information verification methodology, establishing a framework for future development in hybrid neural network architectures focused on complex classification tasks. The validated performance improvements provide strong empirical support for the continued development and implementation of multi-layer classification systems in practical applications requiring sophisticated information authenticity assessment.

Acknowledgements

The authors would like to thank the Armed Forces of Ukraine for the opportunity to write a valid work during the full-scale invasion of the Russian Federation on the territory of Ukraine. Also, the authors wish to extend their gratitude to Kharkiv National University of Radio Electronics for providing licences for additional software to prepare algorithms and the paper.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly Edu and submodule of Microsoft 365 in order to check grammar and spelling. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] I. E. Aïmeur, S. Amri, G. Bassard, Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining 13 (2023) no. 30. doi: 10.1007/s13278-023-01028-5

[2] D. D. Giandomenico, J. Sit, A. Ishizaka, D. Nunan, Fake news, social media and marketing: A systematic review. Journal of Business Research 124 (2021) pp. 329–341. doi: 10.1016/j.jbusres.2020.11.037.
[3] Y. M. Rocha, G. A, de Moura, G. A. Desiderio, C. H. de Oleveira, F. D Lourenco, L. D. de Figueiredo Nicolete, The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. Journal of Public Health 31 (2023) pp 1007–1016. doi: 10.1007/s10389-021-01658-z
[4] A. Khovrat, V. Kobziev, Using Naïve Bayes Classifier to Indentify Falsified Text Information. in: Proceedings of the IEEE 5th KhPI Week on Advanced Technology, Kyiv, Ukraine, 7–10 October 2024, pp. 1–4. doi: 10.1109/KhPIWeek61434.2024.10877950.
[5] A. Khovrat, V. Kobziev, Using RCNN to Indentify the Fake Audio Information. in: Proceedings of the IEEE 7th International Conference on Actual Problems of Unmanned Aerial Vehicles Development, Kyiv, Ukraine, 22–24 October 2024, pp. 205–208. doi: 10.1109/APUAVD64488.2024.10765907.
[6] A. Choudhary, A. Arora, Linguistic feature-based learning model for fake news detection and classification. Expert Systems with Applications. 169 (2021) no. 114171. doi: 10.1016/j.eswa.2020.114171.
[7] S. Garg, D. K. Sharma, Linguistic features-based framework for automatic fake news detection. Computers & Industrial Engineering 172 (2022) no. 108432. doi: 10.1016/j.cie.2022.108432.
[8] N. F. Baarir, A. Djeffal, Fake News detection Using Machine Learning, in: Proceedings of the 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being, Algeria, Algeria, 9–10 February 2021, pp. 205–208. doi: 10.1109/IHSH51661.2021.9378748.
[9] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, J. Vilares, Sentiment Analysis for Fake News Detection. Electronics. 10 (11) (2021) no. 1348. doi: 10.3390/electronics10111348.
[10] S. Kumar, N. A. Jailani, A. R. Singh, S. Panchal S, Sentiment Analysis on Online Reviews using Machine Learning and NLTK, in: Proceedings of the 6th International Conference on Trends in Electronics and Informatics, Tirunelveli, India, 28–30 April 2022, pp. 1183–1189.
[11] M. H. Goldani, R. Safabakhs, S. Momatazi, Convolutional neural network with margin loss for fake news detection. Information Processing & Management. 58 (2021) no. 102418. doi: 10.1016/j.ipm.2020.102418.
[12] M. Chen, Classification with Convolutional Neural Networks in MapReduce. Journal of Computer and Communications. 12 (2024) pp. 174–190. doi: 10.4236/jcc.2024.128011.
[13] P. R. Kanna, P. Santhi, An Enhanced Hybrid Intrusion Detection Using Mapreduce-Optimized Black Widow Convolutional LSTM Neural Networks. Wireless Pers Commun. 138 (2024) pp. 2407–2445. doi: 10.1007/s11277-024-11607-0.