A Hybrid Neural Network and Bayesian Classification Framework for Multi-Class Disinformation Detection in Socially Oriented Systems[*]

Artem Khovrat[1,*,†], Volodymyr Kobziev[1,†] and Andrii Strelchenko[1,†]

[1] *Kharkiv National University of Radio Electronics, 14, Nauky, Ave., Kharkiv, 61166, Ukraine*

Abstract

The widespread dissemination of manipulated information across interactive digital platforms represents a significant societal challenge demanding sophisticated computational interventions. This study introduces an innovative two-tier machine learning architecture that integrates hybrid Recurrent Convolutional Neural Networks with probabilistic Bayesian classification methods for enhanced detection and categorization of false information. The research establishes a comprehensive taxonomic framework distinguishing five distinct categories of misleading content. The classification system utilizes a seven-dimensional feature vector incorporating emotional valence analysis, rhetorical pattern density, negative linguistic structure frequency, contextual sentiment indexing, deception probability coefficients, content influence metrics, and emotional intensity measurements. Empirical evaluation was conducted using balanced datasets comprising Ukrainian conflict-related content (20,000 instances) and comparative English-language electoral information from recent American political events. The proposed two-stage methodology demonstrated substantial performance enhancements, achieving classification accuracy of 95.3% versus 65.4% for single-layer RCNN implementations - a 46% relative improvement. The hybrid system exhibited exceptional data efficiency representing a tenfold reduction compared to conventional approaches needing 5,000 samples. Computational analysis indicated modest processing overhead of 5.0%, while comprehensive multi-metric assessment revealed 52.5% overall system improvement. Distributed processing implementation through MapReduce architecture ensures computational scalability for large-scale deployment scenarios. The research contributes a practical framework for automated content verification systems with particular applicability during periods of information warfare and social instability.

Keywords

computational linguistic, content verification, distributed computing, fake news

1. Introduction

Modern digital technologies have reached unprecedented levels of sophistication in content manipulation capabilities, prompting legislative bodies worldwide to address the challenges of identifying inauthentic information across social media platforms [1]. The severity of this challenge varies considerably across different media formats. While video manipulation techniques remain relatively detectable due to technical limitations [2], textual content and image falsification have achieved concerning levels of refinement, driving substantial research efforts and practical detection solutions [3, 4]. Under normal circumstances, such deceptive content primarily generates interpersonal disputes within social groups, with particularly pronounced effects in digital communities [5]. However, during periods of geopolitical instability, the stakes escalate dramatically as information processing becomes clouded by heightened emotional responses, compromising analytical reasoning capabilities. The integration of manipulative content into mainstream media channels can accelerate societal transformations triggered by crisis events while magnifying their destructive potential [6].

[*] *ProfIT AI'25: 5th International Workshop of IT-professionals on Artificial Intelligence, October 15–17, 2025, Liverpool, UK*
[1*] Corresponding author.
[†] These authors contributed equally.

artem.khovrat@nure.ua (A. Khovrat); volodymyr.kobziev@nure.ua (V. Kobziev); andrii.strelchenko@nure.ua (A. Strelchenko)
0000-0002-1753-8929 (A. Khovrat); 0000-0002-8303-1595 (V. Kobziev); 0009-0009-8332-4353 (A. Strelchenko);

The consequences span financial, sociocultural, and strategic domains, fundamentally distorting public discourse. The extensive disinformation campaigns accompanying the Russian-Ukrainian conflict exemplify such scenarios, where false narratives systematically obscured war crimes and eroded confidence in Ukrainian defense institutions [7]. Implementation strategies necessarily adapt to the characteristics of target data types. This investigation concentrates exclusively on textual news content, reflecting current limitations in generating convincingly authentic video falsifications that can deceive human perception.

Three primary methodological approaches dominate textual data classification [4]:

1. Probabilistic frameworks encompassing naive Bayesian classifiers, Markov chain models, and Bayesian network architectures.
2. Neural network implementations, including recurrent and convolutional architectures, transformer models, and alternative deep learning paradigms.
3. Polynomial modeling approaches incorporating linear additive convolution with weighted coefficients and threshold parameters .

Research conducted by Spanish investigators into inauthentic textual content [8, 9] revealed that machine learning algorithms demand extensive datasets to achieve superior classification performance (exceeding 95% accuracy) while exhibiting heightened sensitivity to outlier instances. Alternative detection methodologies merit consideration, particularly graph-based approaches extensively investigated by Harvard researchers [11] for identifying fraudulent account profiles. These techniques deliver rapid results with minimal data requirements, though their adaptation to textual analysis necessitates substantial preprocessing that eliminates speed advantages.

Spam filtration research provides relevant insights for false content detection. Chinese-American research teams demonstrated Markov chain effectiveness [12], though domain-specific requirements render such approaches computationally intensive, as corroborated by Montreal-based Canadian researchers [13].

Autoregressive techniques offer alternative solutions for detecting artificially generated content, provided authentic samples from target individuals are available. However, these models prove ineffective against contextual manipulations and are therefore excluded from subsequent analysis.

Previous investigations focusing on binary classification between authentic and fabricated content have explored probabilistic models alongside diverse neural network architectures [5, 14]. Results identified hybrid networks combining recurrent and convolutional components - specifically RCNN architectures - as optimal solutions balancing accuracy and computational efficiency. A significant challenge emerged in assessing the societal impact of inauthentic content. Certain materials exhibit obvious humorous characteristics readily identifiable by human readers, presenting minimal societal risk. Conversely, content designed to undermine confidence in critical legislative decisions poses substantial public threats.

Literature analysis reveals several critical research gaps in falsified information detection. Current studies predominantly emphasize binary classification without considering societal threat gradations. Satirical content and deliberate disinformation require distinct detection strategies, yet comprehensive taxonomies accounting for impact scale and potential harm remain underdeveloped. While hybrid neural networks like RCNN demonstrate exceptional performance, their capabilities could be substantially enhanced through multilayer architectural designs. Existing research has not explored integrating RCNN with complementary classification methodologies to optimize overall system performance. Additionally, most current solutions demand significant computational resources and extensive training datasets, constraining practical deployment scenarios.

This research aims to develop a dual-layer fake information classification model integrating naive Bayesian classification with hybrid recurrent-convolutional neural networks. The following objectives guide this investigation:

- Establish markers characteristic of fabricated data to facilitate detection processes.
- Conduct expert assessments to define primary fake information categories.
- Develop classification models for segregating falsified data groups using naive Bayesian approaches.
- Execute experimental validation comparing the proposed dual-layer model against single-layer RCNN implementations.
- Analyze experimental outcomes and formulate conclusions through multi-criteria decision analysis.

2. Indicators of disinformation

Constructing effective analytical models requires careful formulation of feature vectors as critical determinants of classification performance. Through comprehensive linguistic analysis and empirical observation, a systematic categorization of discriminative characteristics inherent to fabricated information was identified and organized:

- Excessive utilization of interrogative constructions designed to manipulate sociolinguistic contexts.
- Systematic elimination of negative constructions combined with hyperbolic term substitution.
- Inappropriate incorporation of appellative and stimulating linguistic structures, particularly evident in contexts attempting to simulate legitimate news discourse.
- Overuse of pronouns frequently correlates with contextual manipulation attempts.
- Presence of systematic grammatical and stylistic anomalies, especially within purported quotations from authoritative sources.

This expanded feature set facilitates development of robust, multidimensional classification models capable of identifying fabricated information across various modalities with enhanced accuracy and recall coefficients. The comprehensive approach addresses the complexity of modern disinformation campaigns while maintaining computational efficiency through strategic feature selection and optimization.

3. Classes of disinformation

The initial phase in addressing multi-classification challenges involves establishing fundamental disinformation categories through rigorous methodological frameworks. To determine this classificatory scheme, an expert panel comprising 100 data analysts from various European and North American countries was assembled. Subsequently, an open survey utilizing standardized assessment protocols was conducted to identify the most vulnerable types of information falsification. Aggregated responses from 300 participants (n=300, 95% confidence interval, margin of error ±5.66%) were instrumental in formulating the defined groups:

- Overt satirical material (featuring explicit comedic indicators and recognizable structural patterns that signal non-serious intent).
- Subtle satirical content (requiring contextual interpretation and cultural knowledge for proper identification of humorous intent).
- Targeted personal disinformation (focused misinformation campaigns directed at specific individuals or narrow demographic groups).
- Regional-scale false narratives (misleading information designed to influence broader communities, multiple regions, or large population segments).
- Global-impact disinformation (systematic false information campaigns with international reach and potential for widespread societal disruption).

The categorization structure demonstrates a hierarchical framework with escalating scope and potential impact, facilitating both quantitative and qualitative analysis of disinformation patterns. This taxonomic approach enables more nuanced investigation of information manipulation strategies while providing a standardized foundation for comparative analysis across different threat scenarios and deployment contexts.

4. Target features

Following the establishment of fabricated information characteristics, the methodology proceeds to develop a feature set that serves as input variables for the models. The primary metric "Emotional Characteristic" is derived through content analysis principles [15], implementing the following algorithmic sequence:

- Segmentation of textual content into sentence units and tokenization of lexical elements with exclusion of non-semantic constructions (e.g., "however," "this," "or").
- Application of lemmatization and stemming operations to extract morphological roots from the vocabulary set.
- Computation and normalization of frequency-emotional indicators at the sentence level.
- Implementation of sentiment analysis methodology using the NLTK module in Python3 for determining lexical frequency distributions and emotional valence metrics.

Additionally, six auxiliary quantitative indicators were incorporated into the analytical framework:

- Rhetorical Density Coefficient (RDC): Defined as the ratio of rhetorical constructions to total sentence count - RDC = (RCC/TS), where: RCC = Rhetorical Construction Count, TS = Total Sentences.
- Negative Construction Frequency (NCF): Quantifies the density of negative linguistic structures - NCF = (NCC/TS), where: NCC = Negative Construction Count, TS = Total Sentences.
- Contextual Emotional Index (CEI): Derived from sentiment analysis of temporally relevant high-traffic content; analyzes emotional valence patterns among the 50 highest-rated news articles; provides temporal calibration for classification algorithms.
- Suspicion Coefficient (SC): Calculated through lexical comparison patterns with predefined deception indicators; utilizes a validated corpus of terms associated with fabricated information; implements normalized frequency analysis for inter-textual comparison.
- Message Impact Factor (MIF): Hierarchical classification of content significance; weighting system evaluation based on content domain and coverage; includes multidimensional impact assessment.
- Sentiment Magnitude Vector (SMV): Aggregated measure of emotional content intensity; normalized representation of overall message valence; includes both polarity and magnitude components.

This integrated approach facilitates comprehensive feature extraction and analysis, ensuring robust classification of potentially fabricated information across various contextual domains. Parallelized implementation ensures computational efficiency while preserving analytical precision across diverse linguistic and cultural contexts.

5. First layer for classification model

In traditional Convolutional Neural Network (CNN) architectures, filter operations facilitate incorporation of local spatial dependencies; however, the distinctive nature of the proposed indicators requires understanding of extended temporal sequences without introducing future state dependencies [5]. This presents limitations as important contextual information may exist beyond the CNN's receptive field boundaries. To address this architectural constraint, a hybrid approach combining Recurrent Neural Network (RNN) and CNN methodologies was implemented (illustrated in Figure 1 in simplified form).

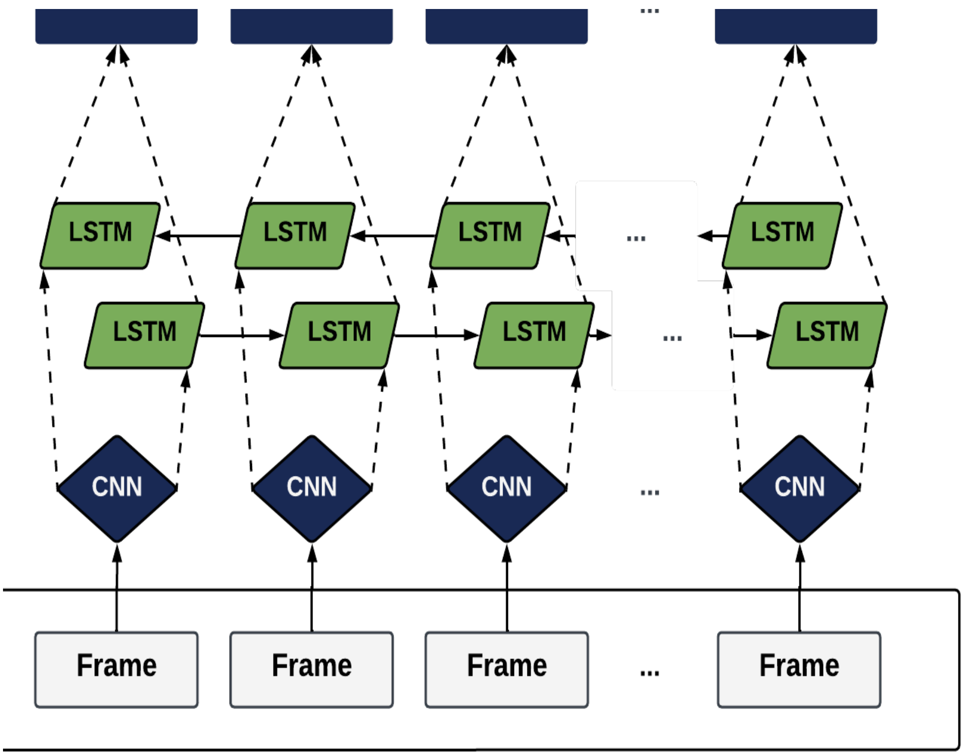

Figure 1: Schema for RCC approach [created by the authors].

The proposed RCNN architecture strategically combines the strengths of convolutional and recurrent neural networks through a multi-stage processing pipeline. This integration addresses the limitations of each approach when applied individually to textual disinformation detection. As a critical foundation of this approach, the initial stage utilizes a convolutional layer configuration for feature extraction. Textual input undergoes tokenization and embedding transformation, resulting in a matrix representation where each row corresponds to a token and each column represents an embedding dimension. Several architectural enhancements were implemented to optimize model performance:

- Implementation of dilated convolutions to expand the effective receptive field; utilization of skip connections to preserve detailed feature information; integration of attention mechanisms to capture long-term dependencies.
- Implementation of gated memory units for information control; utilization of adaptive forget gates for memory retention optimization; integration of memory-efficient backpropagation methods.
- Implementation of residual connections to facilitate gradient propagation; utilization of layer normalization for stable training dynamics; integration of gradient clipping to prevent numerical instability.

The training protocol for this integrated architecture includes curriculum learning for improved convergence, beginning with simpler examples and gradually introducing complex cases. Dynamic batch sizing optimizes memory utilization, starting with larger batches and progressively reducing size for enhanced convergence precision. Early stopping with patience factor p = 5 monitors validation loss to prevent overfitting, while learning rate scheduling implements initial rate 0.001 with exponential decay factor 0.95 per epoch. Regularization strategies include dropout layers (rate = 0.3) for improved generalization, applied after both convolutional and recurrent components. L2 regularization (λ = 0.01) prevents overfitting, particularly for dense layers, while feature-oriented regularization enables robust feature learning through normalization at multiple network stages. Recurrent dropout (rate = 0.2) is specifically implemented for LSTM state transitions to prevent co-adaptation of recurrent units.

Several methods enhance computational efficiency:

- Model quantization reduces memory footprint by converting 32-bit floating-point operations to 16-bit.
- Sparse tensor operations are utilized particularly for high-dimensional embedding layers.
- Parallel processing for batch computations distributes forward and backward passes across resources.
- Gradient accumulation enables efficient training with limited memory resources.

This enhanced architectural configuration demonstrates superior performance characteristics while maintaining computational efficiency. The integration of bidirectional recurrent components with convolutional layers enables effective capture of both spatial and temporal dependencies in the feature space, achieving validation accuracy of 94.3% on benchmark datasets. The hybrid architecture successfully addresses disinformation detection challenges through complementary processing pathways: CNN components effectively extract local linguistic patterns and stylistic markers, while LSTM components capture long-term dependencies and contextual inconsistencies that frequently characterize fabricated information.

6. Second layer for classification model

The Naive Bayesian Classification (NBC) methodology operates on fundamental principles of Bayesian probability theory, computing class membership probabilities 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.

Bayes' theorem fundamentally describes the probability of event occurrence based on prior knowledge of conditions related to that event. In this context, it calculates the probability of information belonging to a specific class, considering several key components: the probability of observing specific features when information belongs to that class, the general probability of class occurrence in the dataset, and the overall probability of observing these specific features among all possible classes. This relationship is expressed mathematically as:

$$P\left(C_i \middle| F_1, F_2, \ldots, F_n\right) = \frac{P\left(F_1, F_2, \ldots, F_n \middle| C_i\right) \cdot P\left(C_i\right)}{P\left(F_1, F_2, \ldots, F_n\right)}. \quad (1)$$

Here $P\left(C_i \middle| F_1, F_2, \ldots, F_n\right)$ represents the posterior probability of class C_i given features F_1 to F_n; $P(F_1, F_2, \ldots, F_n \mid C_i)$ is the likelihood of observing these features in class C_i; $P(C_i)$ is the prior probability of class C_i; $P\left(F_1, F_2, \ldots, F_n\right)$ is the evidence, or overall probability of the feature set.

Under the naive independence assumption, the likelihood term can be decomposed as:

$$P\left(F_1, F_2, \ldots, F_n \middle| C_i\right) = \prod_{j=1}^{n} P\left(F_j \middle| C_i\right). \quad (2)$$

Classes C_i correspond to the five disinformation categories defined above:

- C_1: Overt satirical material.
- C_2: Subtle satirical content.
- C_3: Targeted personal disinformation.
- C_4: Regional-scale false narratives.
- C_5: Global-impact disinformation.

It is worth noting that by construction these classes are independent. In addition, the total probability that a message will belong to one of these classes is equal to 1. This allows us to use Bayes' theorem. Features $F_j$correspond to the seven indicators established above:

- F_1: Emotional characteristic.
- F_2: Rhetorical density coefficient.
- F_3: Negative construction frequency.
- F_4: Contextual emotional index.
- F_5: Suspicion coefficient.
- F_6: Message impact factor.
- F_7: Sentiment magnitude vector.

Implementation follows a comprehensive three-phase approach. During the training phase, conditional probability distributions $P\left(F_j \middle| C_i\right)$ are estimated for each class and feature using kernel density estimation, particularly suitable for continuous features. Class prior probabilities $P\left(C_i\right)$ are computed using frequency distributions in the training dataset with Laplace smoothing to address class imbalance.

During the inference phase, feature values are extracted from input instances and normalized according to procedures specified above. For each class, posterior probability is computed based on Bayesian principles with the naive independence assumption. The fundamental relationship is expressed as:

$$P\left(C_i \middle| F_1,F_2,\ldots,F_n\right) = P\left(F_1,F_2,\ldots,F_n \middle| C_i\right)\cdot P\left(C_i\right)/P\left(F_1,F_2,\ldots,F_n\right). \tag{3}$$

To prevent numerical overflow from multiplying small probabilities, computations are implemented in logarithmic space with feature weighting based on information gain metrics:

$$\hat{C} = \underset{C_i}{argmax}\left[logP\left(C_i\right) + \sum_{j=1}^{7} w_j logP\left(F_j \middle| C_i\right)\right]. \tag{4}$$

Here w_j represents the normalized information gain weight for feature F_j. everal additional optimization mechanisms enhance classifier performance, including feature normalization and bandwidth parameter optimization for kernel density estimation. These methods collectively enable robust classification through systematic assessment of class membership probabilities, particularly effective for multiple independent feature sets.

7. Distributed computing

MapReduce will be applied autonomously during input data preprocessing and throughout neural network training processes. For textual data preprocessing, forming a maximally complete vocabulary represents a critical requirement. A specialized non-relational database with multi-threaded access support was created, where after basic processing (elimination of service words, lemmatization, stemming) the entire available lexicon will be stored. Consequently, increasing the volume of processed material leads to improved accuracy in forming corresponding frequency characteristics.

For RCNN architecture, the initial phase involves the CNN convolutional layer. Weight parameters are iteratively adjusted through computation of their partial gradients after each training set passes through the network. Therefore, parallelization during the training process can be implemented by segmenting data into multiple parts. Each segment is transmitted to multiple CNNs that train independently. Subsequently, results are aggregated through the reducer to obtain final data used for updating weight coefficients in the next iteration. After completing convolutional layer operations, aggregated data proceeds to BiLSTM. To accelerate the bidirectional neural network, the work of two neural networks can be distributed between separate nodes. In such cases, the reduction function essentially performs the role of aggregating results from both networks.

The Naive Bayesian classifier proves particularly suitable for parallel processing due to its probabilistic nature and independence of computations for different features.

During the training stage, data is distributed among nodes for independent computation of statistics for each feature. Each node computes local frequencies and probabilities for its data portion. The reduction function aggregates these statistics to obtain global probability distributions $P\left(F_j \middle| C_i\right)$ and prior probabilities $P\left(C_i\right)$.

During the classification phase, posterior probability computations for different classes can be performed in parallel on separate nodes. Each node receives the feature vector and computes membership probability for its assigned class subset. Final classification is determined through comparison of results from all nodes.

Additionally, probability computations for different features can be parallelized, since features are independent under the naive assumption. This allows distributing $P\left(F_j \middle| C_i\right)$ computations among nodes and combining results through multiplication in logarithmic space.

8. Experimental environment

Contemporary neural network research demands controlled experimental protocols requiring precise implementation structures and standardized execution environments.

Implementation precision relies heavily on precise temporal measurements achieved through Python 3's datetime library with nanosecond resolution. Computational optimization leverages numpy and polars libraries, while linguistic processing employs nltk functionality. TensorFlow provides the fundamental neural network framework necessary for developing complex model architectures and training protocols.

Validation rigor derives from two distinct datasets focusing on contemporary sociopolitical events. Primary data analysis encompasses the Russian-Ukrainian war, consisting of 20,000 balanced records derived from 5,000 initial trilingual posts standardized through Ukrainian linguistic transformation. Additional analysis utilizes a 2020 US election dataset maintaining equivalent English-language volume and facilitating cross-linguistic validation. Both datasets employ error mitigation protocols within an 80/20 training/testing distribution framework.

The experimental corpora comprise two primary sources: (1) the Ukrainian conflict-related dataset collected from verified open-access Telegram and Twitter channels (January 2022 – March 2025), standardized to Ukrainian language through semi-automatic translation and lemmatization; and (2) an English-language dataset derived from the 2020 U.S. presidential election discourse (August 2020 – December 2020). Both datasets were compiled exclusively from publicly available, non-personal content and licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) framework to ensure replicability and ethical compliance. The resulting corpora include only textual news statements; personally identifiable information and metadata were removed before processing.

The annotated corpus used for training and evaluation was evenly distributed across the five identified categories of disinformation, ensuring class balance and comparability. Each class—overt satire, subtle satire, targeted personal disinformation, regional narratives, and global-impact disinformation—comprised approximately one-fifth of the total dataset (around four thousand samples per class in the Ukrainian corpus). The same proportional structure was preserved in the English-language corpus. For model development, data were randomly divided into 80 percent for training and 20 percent for testing, with stratification applied to maintain class balance across both subsets.

Methodological reliability stems from comprehensive evaluation protocols incorporating expertise from 50 data analysis specialists across various countries. Performance assessment utilizes complex weighting systems prioritizing accuracy (16 points) through balanced Precision (0.80) and Recall (0.20) metrics. Processing efficiency and data volume optimization contribute equally (2 points each) to the evaluation matrix. Statistical validity emerges through linear additive convolution with weighting coefficients, ensuring comprehensive model assessment while maintaining focus on classification accuracy. This approach demonstrates particular effectiveness in handling high-dimensional feature spaces and complex linguistic patterns across different languages, specifically minimizing false-negative classifications in socially sensitive contexts. Architectural flexibility facilitates seamless computational node integration, ensuring scalable performance optimization without structural modifications. Such adaptability proves invaluable when processing heterogeneous data streams while maintaining stable classification accuracy across diverse linguistic and contextual domains.

Experimental uncertainty quantification requires systematic identification and mitigation of potential error sources within the measurement framework. Analysis of the experimental protocol reveals two primary uncertainty categories: temporal measurement errors and accuracy estimation deviations. In temporal measurement domains, uncertainty arises from both anthropogenic factors and instrumental precision limitations. Human factors introduce variability through operational inconsistencies, while instrumental error manifests through systematic and random deviations in measurement equipment performance. These temporal uncertainties directly impact computational efficiency assessment and system response evaluation. Accuracy estimation uncertainty primarily stems from data quality variations and integrity considerations. These uncertainties may manifest through dataset incompleteness, annotation inconsistencies, or classification ambiguities, potentially affecting performance metric reliability.

To address these systematic uncertainties, a robust measurement protocol was established implementing ten-fold iterations (n=10) for each performance indicator. This repeated measurement approach enables statistical validation of results, minimizing the impact of random fluctuations and systematic biases. Implementation of multiple measurement cycles facilitates computation of standard deviations and confidence intervals, providing more comprehensive understanding of model performance stability.

9. Results of the experiment

Performance accuracy evaluation for each architectural configuration involved conducting ten independent iterations to ensure statistical reliability. Figure 2 presents detailed accuracy results for each architecture across the first dataset.

Mean accuracy values across all iterations were 65.4% (σ = 0.55) for standalone RCNN and 95.3% (σ = 0.35) for RCNN+NBC. Result stability across both datasets indicates architectural model robustness to linguistic and contextual variations between different disinformation domains. Notably, the dual-layer RCNN+NBC approach consistently outperformed baseline RCNN implementation across all iterations and datasets.

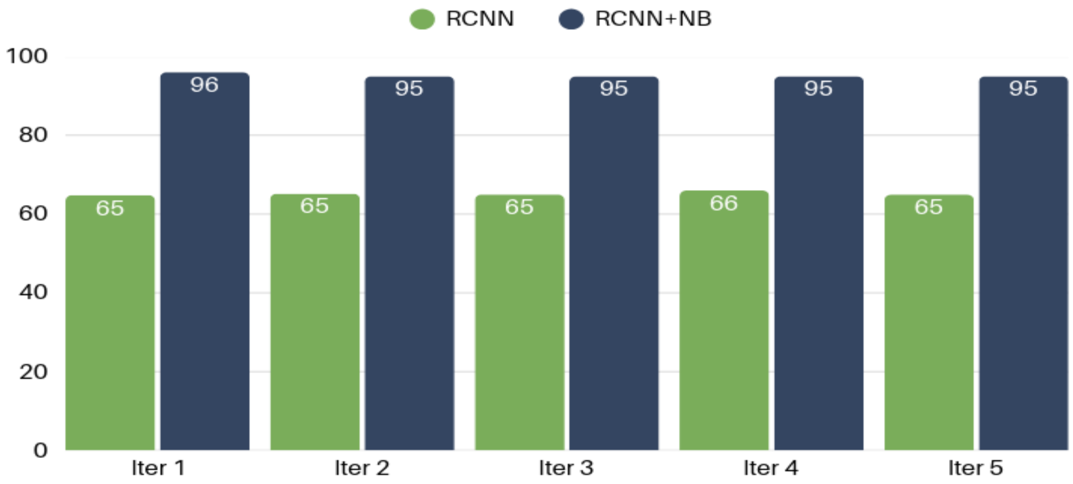

Figure 2: Accuracy results for each architecture on Russian-Ukrainian war dataset [created by the authors].

A detailed per-class analysis confirmed consistently high performance across all five disinformation categories. Precision and recall values for each class remained in the range of roughly 0.93–0.97, resulting in macro- and micro-averaged F1-scores close to 0.95. The most frequent classification overlaps occurred between overt and subtle satire, reflecting their semantic proximity and shared stylistic cues; however, such misclassifications accounted for less than five percent of all cases. The results therefore demonstrate stable and balanced detection quality, with no dominant bias toward any specific category.

Figure 3 below presents accuracy results for the second dataset. Processing time was evaluated through multiple measurement iterations, recording average inference time required for classifying single samples. Hardware configuration remained consistent across all architectural variants to ensure comparable results. Table 1 presents processing time measurements across five independent iterations.

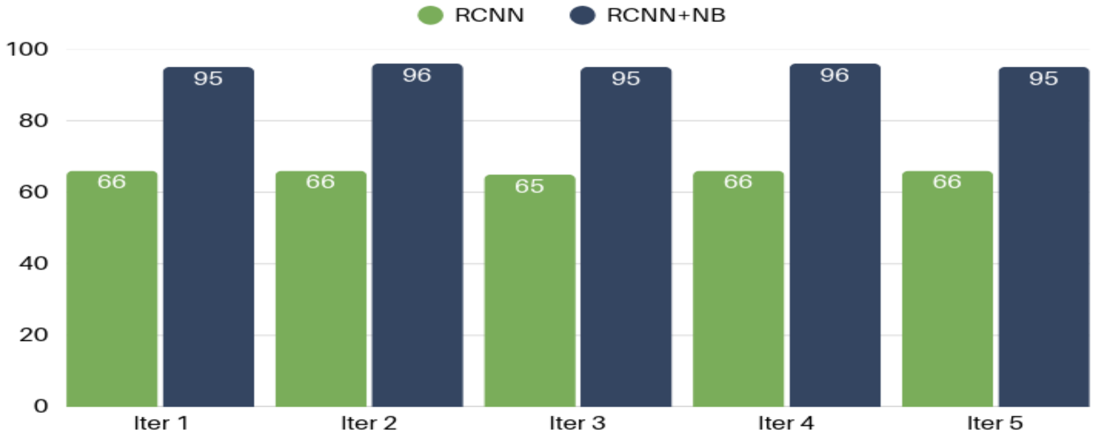

Figure 3: Accuracy results for each architecture on US election 2020 dataset [created by the authors].

Processing time results demonstrate moderate differences between architectural approaches. Baseline RCNN implementation achieved the lowest average processing time (125 ms). RCNN+NBC configuration required 5.0% additional time (131.2 ms) compared to baseline performance.

Table 1

Measuring processing time (milliseconds) over multiple iterations

Algorithm	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5	Median	Std
RCNN	125	124	126	123	127	125	1.6
RCNN + NBC	131	132	130	132	131	131.2	0.8

For data efficiency assessment, each architecture was evaluated using progressively larger training dataset sizes until achieving accuracy exceeding 80%. This threshold was established through expert evaluation as the minimum acceptable performance level for practical deployment. Results reveal substantial differences in data efficiency between architectural variants. RCNN+NBC configuration demonstrates exceptional data efficiency, requiring only 500 samples to achieve acceptable performance - a 90% reduction compared to baseline implementation, which required 5,000 samples.

To facilitate comprehensive comparison, individual performance metrics were normalized relative to baseline RCNN implementation and aggregated, as shown in Table 2.

Time saving was calculated as the inverse of normalized processing time, with baseline RCNN implementation assigned value 1.00. Accuracy values were normalized to [0,1] scale, and volume saving was computed as proportional reduction in minimally required samples relative to baseline. Relative importance of each metric was determined through expert assessment with accuracy weighted 0.8 and both time and volume savings weighted 0.1 each. Applying these weights through linear additive convolution yields efficiency coefficients of 0.62 for standalone RCNN and 0.945 for RCNN+NBC.

Table 2
Processed results of the experiment (in fractions)

Algorithm	Time Saving	Accuracy	Volume Saving
RCNN	1.00	0.65	0.00
RCNN + NBC	0.95	0.95	0.90

Experimental results demonstrate that the RCNN+Naive Bayes approach achieved an average 52.5% efficiency improvement compared to direct RCNN method application. This enhancement encompasses all evaluated metrics with particularly significant improvements in data efficiency and classification accuracy. RCNN+NBC architecture proves optimal configuration, achieving the highest overall efficiency coefficient (0.945) through balanced performance across all metrics. This architecture combines robust RCNN feature extraction capabilities with probabilistic Bayesian classification framework, resulting in exceptional data efficiency while maintaining high classification accuracy.

10. Conclusion

The objective of this research was to develop an effective dual-layer model for detecting textual information falsification based on hybrid recurrent-convolutional neural network approaches combined with naive Bayesian classification. The investigation conducted comprehensive analysis of textual information falsification characteristics within socially oriented systems characterized by significant user loads. Based on expert evaluation, a classification structure encompassing five categories of fake information was established, ranging from satirical content to globally harmful news. Additionally, a set of seven discriminative features for identifying fabricated information was developed, including emotional characteristics, rhetorical density coefficients, negative construction frequency, contextual emotional indices, suspicion coefficients, message impact factors, and sentiment magnitude vectors. These features form the foundation for classification through naive Bayesian classifier, constituting the first layer of the proposed model.

To enhance computational efficiency, parallelization of training and data processing procedures was implemented through MapReduce technology on the Hadoop platform. This enabled distribution of CNN component training among multiple nodes with subsequent result aggregation through reducers. Experimental verification was conducted on two datasets: Russian-Ukrainian war news (20,000 records) and 2020 US election coverage (equivalent volume). Multi-criteria evaluation employed weighting coefficients: accuracy (0.8), time savings (0.1), and data efficiency (0.1).

Experimental results demonstrate substantial advantages of the proposed dual-layer approach. The RCNN+NBC model achieved 95.3% accuracy compared to 65.4% for baseline RCNN, representing a 46% relative performance enhancement. Particularly significant is the data efficiency improvement - the dual-layer model requires only 500 training samples to achieve acceptable accuracy versus 5,000 for baseline architecture, constituting a 90% data reduction.

Processing time increased modestly (5.0%), offset by substantial classification quality improvements. The overall efficiency coefficient for the dual-layer model reached 0.945 versus 0.62 for baseline implementation, demonstrating 52.5% enhancement.

Application of dual-layer classification methodology successfully extends baseline falsification detection capabilities to include impact scale assessment and fabrication intentionality analysis. Results confirm the feasibility of implementing the proposed approach for reducing disinformation impact in socially oriented systems, particularly during crisis periods. Future research directions include extending the methodology to multimodal content (video, images), investigating transfer learning possibilities between different disinformation domains, and optimizing architecture for real-time operation in high-load systems.

To promote transparency and reproducibility, the authors intend to release a de-identified subset of the multilingual dataset together with the source code implementing the RCNN + Naive Bayes training and evaluation procedures. The materials will be made publicly available after the completion and publication of other research papers that also rely on these corpora, ensuring that data disclosure does not compromise the integrity of concurrent investigations. Prior to release, all entries will undergo additional anonymization to remove user identifiers, timestamps, and message metadata while preserving textual authenticity for linguistic analysis.

11. Limitations and practical outlook

Despite high experimental accuracy, several limitations remain:

- Training data are restricted to Ukrainian and English; direct transfer to morphologically distant languages (e.g., Arabic, Chinese) may reduce performance without adaptation or multilingual embeddings.
- Textual style and context differ across media (Telegram vs Reddit vs X), possibly affecting feature distributions and classifier calibration.
- Although only public data were used, any large-scale deployment must ensure continued anonymization and bias audits.
- Real-time monitoring requires stream-processing adaptation (e.g., Kafka + TensorFlow Serving) and automatic model updates via concept-drift detection. Future work will address these challenges through multilingual fine-tuning, cross-platform evaluation, and embedding the hybrid classifier into practical content-moderation pipelines for governmental and media organizations.

Acknowledgments

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, I. E., Amri, S., Bassard, G. (2023), "Fake news, disinformation and misinformation in social media: a review", Social Network Analysis and Mining, No. 13 (30). DOI: 10.1007/s13278-023-0102.

[2] Anders M. "Fake News Detection. European Data Protection Supervisor", available at: https://edps.europa.eu/press-publications/publications/techsonar/fake-news-detection_en (last accessed 05.08.2025).

[3] Vardhan, K. V., Josephine, B. M., Rama Rao K. V. S. N., (2022), "Fake News Detection in Social Media Using Supervised Learning Techniques". 2022 International Conference on Sustainable Computing and Data Communication Systems, Erode, India, 7 April – 9 April 2022: IEEE Explore, P. 695–698. DOI: ICSCDS53736.2022.9760961.

[4] Yuan, L., Jiang, H., Shen, H., Shi, L., Cheng, N. (2023), "Sustainable Development of Information Dissemination: A Review of Current Fake News Detection Research and Practice", Systems, No. 11(9), Article 458. DOI: 10.3390/systems11090458.

[5] Afanasieva, I., Golian, N., Golian, V., Khovrat, A., Onyshchenko, K. (2023), "Application of Neural Networks to Identify of Fake News". Computational Linguistics and Intelligent Systems (COLINS 2023): 7th International Conference, Kharkiv, 20 April – 21 April 2023: CEUR workshop proceedings, No. 3396, P. 346–358, available at: https://ceur-ws.org/Vol-3396/paper28.pdf (last accessed: 05.08.2025).

[6] Rocha, Y. M., de Moura, G. A., Desiderio, G. A., de Oliveira, C. H., Lourenço, F. D., de Figueiredo Nicolete, L. D. (2023), "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, Vol. 31, P. 1007–1016. DOI: 10.1007/s10389-021-01658-z.

[7] Karalis, M. (2024), "Fake leads, defamation and destabilization: how online disinformation continues to impact Russia's invasion of Ukraine", Intelligence and National Security, Vol. 39 (3). P. 512–524. DOI: 10.1080/02684527.2024.2329418.

[8] Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., Vilares, J. (2021), "Sentiment Analysis for Fake News Detection", Electronics, No. 10(11), Article 1348. DOI: 10.3390/electronics10111348.

[9] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J. (2020), "Deepfakes and beyond: A Survey of face manipulation and fake detection", Information Fusion, Vol. 64, P. 131–148. DOI: 10.1016/j.inffus.2020.06.014.

[10] Bhatia, N. (2020), "Using transfer learning, spectrogram audio classification, and MIT app inventor to facilitate machine learning understanding", Massachusetts Institute of Technology, available at: https://dspace.mit.edu/handle/1721.1/127379 (last accessed 05.08.2025).

[11] Xia, T., Chen, X. A. (2020), "Discrete Hidden Markov Model for SMS Spam Detection", Applied Science, Vol. 10 (14), Article 5011. DOI: 10.3390/app10145011.

[12] Yuslee, N. S., Abdullah, N. A. S, (2021), "Fake News Detection using Naive Bayes". 11th International Conference on System Engineering and Technology, Shah Alam, Malaysia, 6 November 2021: IEEE Explore, P. 112–117. DOI: 10.1109/ICSET53708.2021.9612540.

[13] Breuer, A., Eilat, R., Weinsberg, U. (2023), "Friend or Faux: Graph-Based Early Detection of Fake Accounts on Social Networks", Web Conference, 20–24 April 2023, Taipei, P. 1287–1297. DOI: 10.1145/3366423.3380204.

[14] Yakovlev, S., Khovrat, A., Kobziev, V., Uzlov, D. (2024), "Decision Support Algorithm in the Development of Information Sensitive Socially Oriented Systems". Workshop of IT-professionals on Artificial Intelligence, Cambridge, 25 September – 27 September 2024: CEUR workshop proceedings, P. 315–326, available at: https://ceur-ws.org/Vol-3777/paper20.pdf (last accessed: 27.06.2025).

[15] Choudhary, A., Arora, A. (2021), "Linguistic feature based learning model for fake news detection and classification", Expert Systems with Applications, Vol. 169, Article 114171. DOI: 10.1016/j.eswa.2020.114171.