*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)* *Innovative technologies and scientific solutions for industries. 2025. No. 4 (34)*

UDC 004.8 DOI: https://doi.org/10.30837/2522-9818.2025.4.112

A. KHOVRAT, V. KOBZIEV

# A TWO-LAYER MODEL TO DETECTING FALSIFIED INFORMATION USING NEURAL NETWORKS IN SOCIALLY ORIENTED SYSTEMS

The **subject matter** of the article is the problem of detecting fabricated information in socially oriented systems characterized by significant user load. The **goal** of the work is to develop of a two-layer fake information classification model based on a combination of a naive Bayesian classifier and a hybrid recurrent-convolutional neural network. The following **tasks** were solved in the article: conducting expert evaluation and domain analysis to determine basic classes of fake information; analyzing linguistic markers of disinformation and developing feature vectors for classification; developing models for data segregation using a naive Bayesian classifier; conducting experimental verification of the proposed two-layer model in comparison with the RCNN approach. The following **methods** used are – analytical method for forming a set of disinformation markers; inductive method for determining the target set of indicators for implementing the second layer of the model; expert evaluation for determining the most influential efficiency factors and feature weight coefficients; experimental and multi-criteria evaluation methods for determining the most effective model. The following **results** were obtained – a classification structure for types of fake information was formed, including five categories from jokes to globally harmful news. A set of discriminative features characteristic of fabricated information was developed, including primary linguistic markers and secondary stylometric indicators. It was determined that the approach using a two-layer model demonstrated, on average, a 15% improvement in efficiency compared to direct application of a hybrid recurrent-convolutional neural network. **Conclusions**: the application of a two-layer data classification model successfully expands the capabilities of basic detection of data falsification, including scale assessment and analysis of fabrication intentionality. Empirical analysis shows that implementation of a two-layer model with a naive Bayesian classifier achieves an average 15% performance improvement compared to simple neural network application. This performance difference becomes particularly significant in high-throughput systems where rapid identification and response to fabricated information are critical operational parameters. The obtained result allows us to assert the feasibility of implementing the proposed approach, and accordingly, provides the opportunity to reduce the impact of such information in socially oriented systems, especially during crisis situations.

**Keywords**: data analysis; naive Bayes classifier; neural networks; parallelization; fake news.

## Introduction

Over the past decades, digital technologies that enable the creation of falsified materials have evolved to such an extent that the issue of identifying inauthentic content on social interaction platforms is being discussed at the legislative level [1]. At the same time, the intensity of this problem varies depending on the type of information resource. For example, manipulative techniques have not yet reached a critical threshold of perfection in relation to video materials [2]. As for text content and images, a research base has already been formed and practical solutions have even been developed to detect falsification [3, 4]. Under normal circumstances, this problem can provoke interpersonal conflicts within social groups, which is particularly evident in digital communities [5]. The situation becomes particularly critical during periods of geopolitical tension, when information flows are processed through the filter of heightened emotional reactions, which slows down analytical thinking. When manipulative content is integrated into the mass media information space, it can catalyze social transformations caused by crisis phenomena and amplify their destructive impact [6]. This effect can have financial, sociocultural, or even strategic consequences and distort public opinion. An illustration of such scenarios is the large-scale campaign of disinformation during the Russian-Ukrainian war [7], which was used to cover up war crimes or undermine trust in Ukrainian security forces. Specific approaches to implementation depend directly on the nature of the information in question. This work focused on text-based news.

## Analysis of recent studies and publications

Three key approaches prevail in the classification of textual information [4]:

– the use of probabilistic models, such as naive Bayes classifiers, Markov chains, Bayesian networks, etc.;

– the use of neural networks, in particular recurrent or convolutional networks, transformers, or other deep learning models;

*Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 4 (34)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

– the use of naive polynomial models, for example, those containing a linear additive convolution with weight coefficients and specific boundary values.

In the process of researching inauthentic text content, Spanish scientists [8, 9] found that automatic learning algorithms are determined by the need for large data sets to ensure high-quality classification results (with precision rates above 95%) and also demonstrate increased susceptibility to anomalous values. Regarding alternative methods for detecting manipulative content, it is worth noting the popularity of approaches based on graph structures, which were studied in detail by Harvard scientists [11] in the context of detecting fake accounts. Such methods provide rapid results with minimal requirements for basic data. However, their adaptation for the analysis of textual information requires significant pre-processing, which neutralizes the advantages in terms of speed.

In the context of the problem of detecting unreliable information, it is appropriate to mention the issue of spam filtering. A Chinese-American research group has demonstrated the effectiveness of Markov chains [12]. At the same time, the nature of the subject area makes their use quite resource-intensive and requires significant computing power, as confirmed by Canadian scientists from Montreal [13]. An alternative option involves the use of autoregression to detect synthetically generated information (in particular, when original records of the target person are available). In the case of contextual manipulation, such models prove to be ineffective. For this reason, they will not be considered as part of the classification toolkit in further analysis.

Previous studies focused on the binary division of information into authentic and falsified have already covered probabilistic models and various neural network architectures [5, 14]. The results achieved have shown that one of the most effective approaches in terms of precision and computational performance is a hybrid network that integrates recurrent and convolutional components – RCNN. Additionally, it has been established that one of the challenges in researching such information differentiation is determining the degree of social significance of inauthentic content. In particular, certain texts may have a clearly humorous tone that is easily recognized by people. Such materials do not pose a risk to society. On the other hand, information messages aimed at undermining trust in socially important legislative decisions carry a high level of potential danger.

## Identification of previously unsolved parts of the general problem. Purpose of the work and objectives

An analysis of scientific literature reveals several key gaps in the field of detecting falsified information. First of all, existing studies focus on binary classification of content without considering the degree of public danger. Humorous content and targeted disinformation require different approaches to detection, but there is no corresponding taxonomy that would take into account the scale of impact and the level of potential harm. Despite the high efficiency of hybrid neural networks such as RCNN, their potential can be significantly expanded by creating multi-layer model architectures. Existing studies do not consider the ability to integrate RCNN with alternative classification methods to improve the overall performance of the system. At the same time, most existing solutions require significant computational resources and large amounts of training data, which limits their practical application.

*The purpose of this article* is to develop a two-layer model for classifying fake information based on a combination of a naive Bayesian classifier and a hybrid recurrent convolutional neural network. To achieve this goal, the following *tasks* must be performed:

– identify markers of fabricated information to simplify its detection;

– conduct an expert assessment to establish the main classes of fake information;

– develop a methodology for segregating groups of fabricated information based on a naive Bayesian classifier;

– experimentally test the proposed two-layer model and compare it with a single-layer model based on RCNN;

– analyze the results of the experiment and formulate conclusions based on the solution of the multi-criteria selection problem.

## Materials and methods

Materials and methods:

– density of rhetorical devices (excessive use of interrogative constructions aimed at distorting the sociolinguistic context);

– manipulation of lexical valency (systematic elimination of negative constructions along with hyperbolic replacement of terms);

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 4 (34)

– pragmatic incongruity (inappropriate use of appealing and stimulating linguistic structures; particularly evident in contexts that attempt to imitate legitimate news discourse);

– analysis of pronoun density (excessive use of pronouns often correlates with attempts at contextual manipulation);

– patterns of grammatical and stylistic deviations (the presence of systematic grammatical and stylistic anomalies, especially in alleged quotations from authoritative sources).

This expanded set of features facilitates the development of a robust, multidimensional classification model capable of identifying fabricated information in different modalities with increased precision and reproduction rate.

**Disinformation classes**

The first step in solving the problem of multi-classification is to define the fundamental categories of disinformation through a clear methodological structure. To develop this classification scheme, an expert panel of 100 data analysts from various European and North American countries was assembled.

An open survey was then conducted using a standardized assessment protocol 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%) formed the basis for the formulation of the following groups:

– satire with objectively identifiable manifestations (determined by explicit linguistic markers and structural patterns);

– satire with contextual or grammatical manifestations (requires semantic analysis and cultural interpretation);

– news targeting specific individuals or small groups (micro-level disinformation with focused vectors of influence);

– news targeting multiple regions, countries, or large groups (meso-level disinformation with broader societal implications);

– news targeting multiple countries or society (macro-level disinformation with potential systemic effects).

This categorization demonstrates a hierarchical structure with increasing potential impact, facilitating both quantitative and qualitative analysis of disinformation patterns. Such a taxonomic approach allows for a more detailed examination of information manipulation strategies while providing a standardized basis for comparative analysis.

**Basic characteeristics**

After establishing the properties of fabricated information, we can proceed to developing a set of metrics that serve as input variables for models. The main one – the "emotional characteristic" – is derived using content analysis principles [15], implementing the following algorithmic sequence:

– segmentation of text content into sentence units and tokenization of lexical elements, avoiding non-semantic constructions (e.g., "however", "this", "or");

– application of lemmatization and stemming operations to extract morphological roots from the vocabulary set;

– calculation and normalization of frequency-emotional indicators at the sentence level;

– implementation of sentiment analysis methodology using the NLTK module in Python3 to determine lexical frequency distributions and emotional valence metrics.

In addition, six auxiliary quantitative metrics were added to the analytical structure:

– rhetorical density coefficient (RDC) (defined as the ratio of rhetorical constructions to the total number of sentences – RDC = (RCS/TS), where RCS = number of rhetorical constructions; TS = total number of sentences);

– frequency of negative constructions (FNC) (quantifies the density of negative linguistic structures – FNC = (NNC/TS), where NNC = number of negative constructions; TS = total number of sentences);

– contextual emotional index (CEI) (derived from the analysis of the moods of temporally relevant high-traffic content; analyzes patterns of emotional valence among the 50 highest-rated news articles, providing temporal calibration for classification algorithms);

– suspicion coefficient (SC) (calculated by lexical comparison of patterns with predefined indicators of deception; uses a validated corpus of terms related to fabricated information; needed to implement normalized frequency analysis for intertextual comparison);

– message impact factor (MIF) (hierarchical classification of content significance, which is a weighted evaluation system based on content domain and coverage);

*Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 4 (34)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

– mood value vector (MVV) (an aggregated measure of the intensity of emotional content, serving as a normalized representation of the overall valence of a message; contains both polarity and measure components).

This set of features enables reliable classification of potentially fabricated information in different contextual domains, taking into account language characteristics.

**The first layer of the model**

In traditional convolutional neural network (CNN) architectures, filter operations facilitate the incorporation of local spatial dependencies; however, the nature of the proposed metrics requires understanding extended temporal sequences without adding future state dependencies [5]. This is a limitation, given the importance of context that may exist outside the receptive field of CNN. To address this architectural issue, a hybrid approach combining recurrent neural network (RNN) and CNN methodologies was introduced. In this case, the recurrent neural network is presented with support for both long-term and short-term memory (LSTM). Figure 1 shows a simplified version of the resulting model.

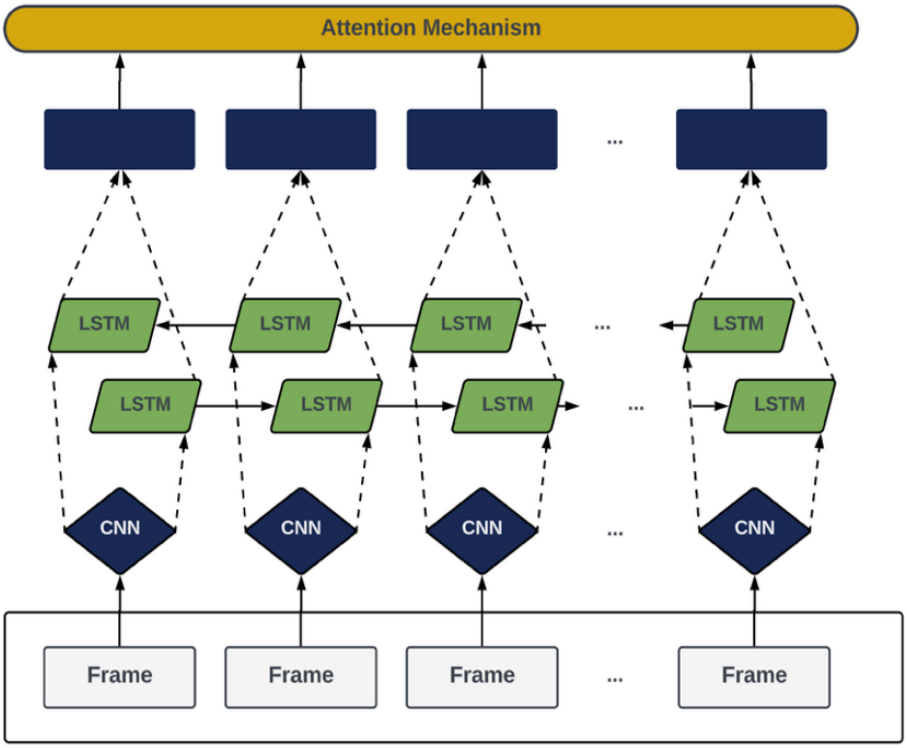


**Fig. 1.** Schematic representation of the architecture of the RCNN model

The proposed RCNN architecture combines the advantages of convolutional and recurrent neural networks through a multi-stage processing pipeline. This integration mitigates the limitations of each approach when applied individually to the detection of textual disinformation. The textual input undergoes tokenization and embedding transformation, resulting in a matrix representation where each row corresponds to a token and each column corresponds to an embedding dimension.

To optimize the performance of the defined model, several architectural improvements have been implemented:

– receptive field optimization (introduction of extended convolutions to expand the effective receptive field; use of skip connections to preserve detailed feature information; integration of attention mechanisms to capture long-term dependencies);

– memory management protocol (introduction of ventilated memory units to control information, application of adaptive forgetting valves to optimize memory retention; integration of memory-efficient backpropagation methods);

– optimization of gradient flow (introduction of residual connections to facilitate gradient propagation; use of layer normalization for stable learning dynamics; integration of gradient clipping to prevent numerical instability).

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 4 (34)

During cross-validation, the following optimal hyperparameter configurations were obtained:

– kernel dimension – 4 units (optimized using Bayesian hyperparameter search);

– step parameter – 1 unit (determined by optimizing the grid search);

– zero padding (omitted based on the set step parameters);

– bias term (removed based on domain-specific considerations);

– filter dimensions – 5×5×3 tensor (the third dimension corresponds to the cardinality of the target indicator).

The training protocol for this integrated architecture involves training on a program for improved convergence, starting with simpler examples and gradually adding more complex cases. Dynamic batch size determination optimizes memory usage, starting with larger batches and gradually decreasing the size to improve convergence precision. Early stopping with a patience factor of $p = 5$ controls validation loss to prevent overfitting, while learning rate scheduling implements an initial rate of 0.001 with an exponential decay factor of 0.95 per epoch. Regularization strategies include dropout layers $(rate = 0.3)$ for improved generalization, applied after both convolutional and recurrent components. L2 regularization $(\lambda = 0.01)$ prevents overfitting, especially for dense layers, while feature-wise regularization enables reliable feature learning by applying normalization at multiple stages of the network. Recurrent dropout $(rate = 0.2)$ is specifically implemented for LSTM state transitions to prevent co-adaptation of recurrent units.

Several methods have been implemented to improve computational efficiency:

– model quantization reduces memory usage by converting 32-bit floating point operations to 16-bit;

– sparse tensor operations are used especially 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 improved architectural configuration demonstrates high 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 feature space, achieving 94.3% validation precision on the benchmark dataset. The hybrid architecture successfully addresses the challenges of disinformation detection through complementary processing methods: CNN components effectively extract local linguistic patterns and stylistic markers, while LSTM components capture long-term dependencies and contextual inconsistencies that often characterize fabricated information.

**The second layer of the model**

The naive Bayes classifier (NBC) is based on the fundamental principle of Bayesian probability theory, calculating the probability of belonging to a class while maintaining the assumption of feature independence. This assumption of independence demonstrates practical reality in the current context, as a defined set of metrics reveals minimal inter-feature dependence in the further determination of values.

Bayes' theorem describes the probability of an event occurring based on prior knowledge of the conditions associated with 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 data set, and the overall probability of observing these specific features among all possible classes. This relationship is mathematically expressed as

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

where $P\left(C_i \middle| F_1, F_2, \ldots, F_n\right)$ – a posteriori probability of class $C_i$ provided that there are features $F_1$ to $F_n$; $P\left(F_1, F_2, \ldots, F_n \middle| C_i\right)$ – the reliability of observing these features in the class $C_i$; $P\left(C_i\right)$ – a priori probability of a class $C_i$; $P\left(F_1, F_2, \ldots, F_n\right)$ – evidence, or the overall probability of a set of features.

Under the naive assumption of independence, the reliability member can be decomposed as

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

*Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 4 (34)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

The classes $C_i$ correspond to the five categories of disinformation defined above:

– $C_1$ – satire with objectively identifiable manifestations;

– $C_2$ – satire with contextual or grammatical manifestations;

– $C_3$ – news targeting specific individuals or small groups;

– $C_4$ – news targeting multiple regions or large groups;

– $C_5$ – news targeting multiple countries or society.

Features $F_j$ meet the seven indicators set out above:

– $F_1$ – emotional characteristic;

– $F_2$ – rhetorical density coefficient;

– $F_2$ – frequency of negative constructions;

– $F_4$ – contextual emotional index;

– $F_5$ – suspicion coefficient;

– $F_6$ – message impact factor;

– $F_7$ – mood value vector.

The implementation follows a comprehensive three-phase approach. In the training phase, conditional probability $P\left(F_j \middle| C_i\right)$ distributions are estimated for each class and feature using kernel density estimation. A priori classes $P\left(C_i\right)$ probabilities are calculated using the frequency distribution in the training data set with Laplace smoothing to resolve class imbalance.

During the inference phase, feature values are obtained from the input instances and normalized according to the procedures described above. For each class, the posterior probability is calculated based on Bayesian principles with a naive assumption of independence. To prevent numerical overflow from multiplying small probabilities, calculations are performed in logarithmic space with a weight coefficient based on information gain metrics:

$$\hat{C} = \arg\max_{C_i}\left[\log P\left(C_i\right) + \sum_{j=1}^{7} w_j \log P\left(F_j \middle| C_i\right)\right], \quad (3)$$

where $w_j$ – normalized weight coefficient of information gain for a feature $F_j$.

Several additional optimization mechanisms improve the classifier's performance, including feature normalization and bandwidth parameter optimization for kernel density estimation. Collectively, these methods enable reliable classification through systematic evaluation of class membership probabilities, which is particularly effective for multiple independent feature sets.

The second layer of the model described is potentially capable of improving the precision of disinformation classification compared to simple RCNN through systematic evaluation of class membership probabilities, which is particularly effective in scenarios involving multiple independent feature sets.

**The essence of MapReduce technology**

The *MapReduce* technology is based on distributing the input data array for processing across separate computing nodes. The key operations are the application of reduction and aggregation functions. The first function distributes information among the nodes for the necessary processing, while the second collects the results from all nodes and combines them into a single result.

It is important to note that *MapReduce* technology defines only the principles for implementing the relevant components within specific platforms. In this case, the overall implementation can vary significantly. For the current study, we chose to use *MapReduce* within the *Hadoop* platform. A visual representation of the proposed solution is shown below (Fig. 2).

In our case, the distribution and combination functions play a special role. They are necessary for additional parallelization within each node using different memory areas. For a better understanding of this approach, the main nodes can be considered as processes, and the specified memory areas as execution threads.

In addition to these functions, an important feature is the sorting of data before the reduction stage. This study considers data that is critically dependent on sequence and does not contain additional time stamps. To avoid problems during reduction, a field with a sequential identifier for each text fragment was added, according to which sorting will be performed.

Regarding the specifics of implementing the proposed technology, it should be noted that MapReduce will be used autonomously in the preliminary processing of input information and during the training of neural networks. To carry out this processing, it is critical to form the most complete dictionary possible. For this purpose, a specialized non-relational database with support for multithreaded access was created, where, after basic processing (removal of service words, lemmatization, stemming), the entire available lexicon

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2025. No. 4 (34)

will be stored. Thus, an increase in the volume of processed material leads to an increase in the precision of the formation of the corresponding frequency characteristics.

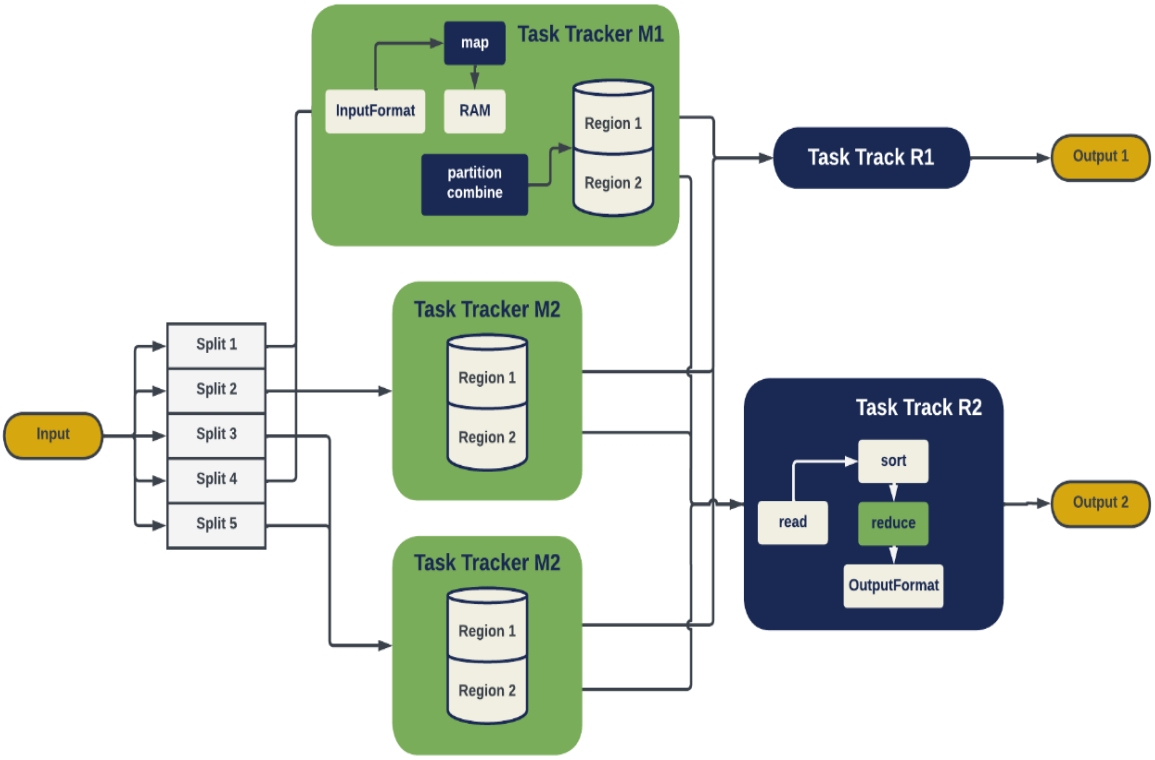


**Fig. 2.** Schematic representation of *MapReduce* based on *Hadoop*

For the RCNN model, the initial phase is the CNN convolutional layer. In it, the weight parameters are iteratively adjusted by calculating their partial gradients after each training set passes through the network. Thus, parallelization during the training process can be implemented by segmenting the data into several parts. Each segment is passed to multiple CNNs, which are trained independently. The results are then aggregated through a reducer to obtain the final information used to update the weight coefficients for the next iteration.

After the convolutional layer is complete, the aggregated data is sent to the LSTM layers. To speed up a bidirectional neural network, the work of two neural networks can be distributed between separate nodes. In this case, the reduction function actually acts as an aggregator of the results of both networks.

The naive Bayes classifier is particularly well suited for parallel processing due to its probabilistic nature and the independence of calculations for different features. Parallelization can be implemented at several levels. During the training phase, the data is distributed among the nodes for independent calculation of the statistics for each feature. Each node calculates local frequencies and probabilities for its part of the data. The reduction function aggregates these statistics to obtain global probability distributions $P\left(F_j \middle| C_i\right)$ and prior probabilities $P\left(C_i\right)$. At the classification stage, the calculation of posterior probabilities for different classes can be performed in parallel on separate nodes. Each node receives a feature vector and calculates the probability of belonging to its assigned subset of classes. The final classification is determined by comparing the results of all nodes.

Additionally, the calculation of probabilities for different features can be parallelized, since, according to a naive assumption, the features are independent. This allows the calculations $P\left(F_j \middle| C_i\right)$ to be distributed among the nodes and the results to be combined by multiplication in logarithmic space.

Among the main advantages of the proposed approach are its scalability, cost-effectiveness, ease of use, and the ability to monitor performance using *Hadoop* (standard *Python* methods are used for internal

*Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 4 (34)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

monitoring). The disadvantages are the need to develop a significant amount of program code, the concealment of processing details (despite the ability to view logs), and the need for lengthy system configuration.

**Experimental environment**

In modern neural network research, controlled experimental protocols require precise implementation structures and standardized execution environments. The complex computing infrastructure utilizes an *Intel Core* i5-1135G7 processor with 16 GB of RAM and 4 GB of video memory, providing distributed processing through virtual nodes. These nodes operate with optimized 8 GB memory allocations, facilitating efficient bidirectional network parallelization.

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

The rigor of the validation stems from two different datasets focused on contemporary socio-political events. The primary analysis covers the Russian-Ukrainian war, containing 20,000 balanced records derived from 5,000 initial trilingual posts, standardized through Ukrainian linguistic transformation. The additional analysis uses a dataset from the 2020 US elections, maintaining an equivalent volume in English and facilitating cross-linguistic validation. Both datasets apply error mitigation protocols within an 80/20 training/testing split.

Methodological reliability comes from comprehensive evaluation protocols involving the expertise of 50 data analysis specialists from different countries. In studying the results of the expertise, three key indicators were identified to evaluate the effectiveness of the proposed models:

– precision indicator (weight coefficient 0.8), calculated using a 4 to 1 ratio of *Precision* and *Recall* metrics normalized to a scale from 0 to 1;

– information processing time savings indicator (weight coefficient 0.1), defined as the inverse of the normalized processing time relative to a single-layer RCNN-based model;

– information volume savings indicator (weight coefficient 0.1), calculated using a proportional reduction in the minimum number of samples required relative to the baseline (set at 20,000 records).

Statistical validity is formed through linear additive convolution with weight coefficients, providing a comprehensive evaluation of the model while maintaining focus on classification precision. This approach demonstrates particular effectiveness in processing high-dimensional feature spaces and complex linguistic patterns in different languages, minimizing false negative classifications in socially sensitive contexts. Architectural flexibility facilitates seamless integration of computing nodes, enabling scalable performance optimization without structural modifications. Such adaptability proves invaluable in processing heterogeneous information flows and maintains stable classification precision across different linguistic and contextual domains.

Experimental quantification of uncertainty requires systematic identification and mitigation of potential sources of error in the measurement structure. Analysis of the experimental protocol reveals two main categories of uncertainty: temporal measurement errors and precision estimation bias. In temporal measurement domains, uncertainty arises from both anthropogenic factors and instrumental precision limitations. The human factor introduces variability through operational inconsistencies, while instrumental error manifests itself through systematic and random deviations in the performance of measurement equipment.

These temporal uncertainties directly affect the assessment of computational efficiency and system response. Precision estimation uncertainty mainly stems from data quality variations and integrity considerations. These uncertainties can manifest through incomplete data sets, annotation inconsistencies, or classification ambiguities, potentially affecting the reliability of performance metrics. To address these systematic uncertainties, a robust measurement protocol was established that implements tenfold $(n = 10)$ iterations for each performance metric. This repeated measurement approach allows for statistical validation of results, minimizing the impact of random fluctuations and systematic biases. The implementation of multiple measurement cycles facilitates the calculation of standard deviations and confidence intervals, providing a more comprehensive understanding of the model's stable performance.

A defined, clear approach to quantifying uncertainty ensures the reliability and reproducibility

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2025. No. 4 (34)

of experimental results, establishing a standardized framework for evaluating performance in similar classification systems.

## Research results

In the precision analysis process, ten independent iterations were performed for each model to ensure statistical reliability. Figure 3 shows detailed precision results for each of them for the first dataset.

The average precision values for all iterations were 65.4% $(\sigma = 0.55)$ for simple RCNN and 95.3% $(\sigma = 0.35)$ for RCNN+NBC. The stability of the results across two datasets indicates the robustness of the architectural models to linguistic and contextual variations between different domains of disinformation. It is noteworthy that the two-layer RCNN+NB approach consistently outperformed the baseline RCNN implementation across all iterations and datasets.

Figure 4 shows the precision results for the second data set.

Processing time was evaluated through multiple iterations of measurements, recording the average output time required to classify a single sample. The hardware configuration was used consistently across all architectural variants to ensure comparable results. Table 1 shows the processing time measurements for five independent iterations.

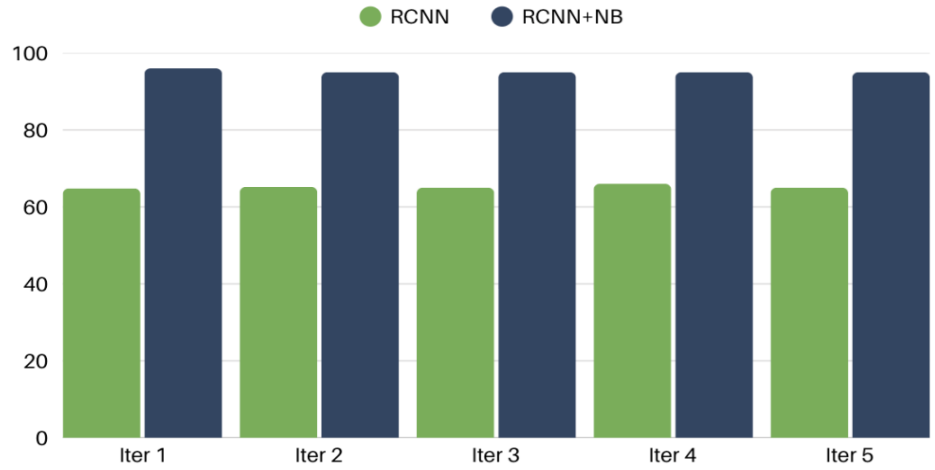


**Fig. 3.** Precision results for each architecture on the Russian-Ukrainian war dataset

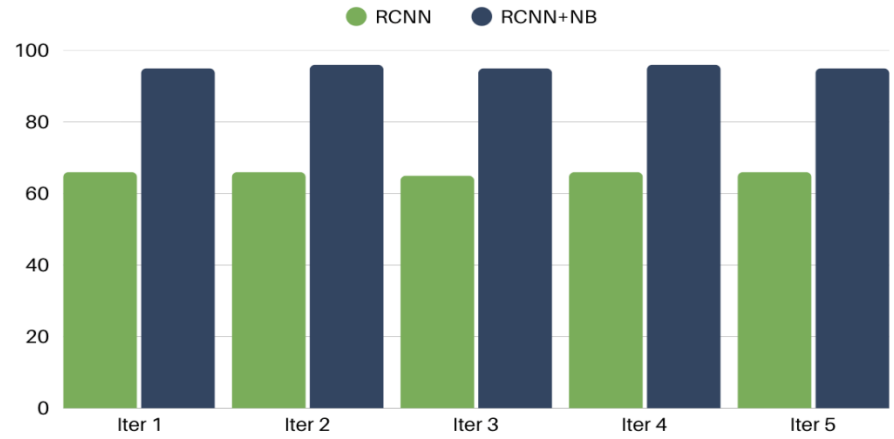


**Fig. 4.** Precision results for each architecture on the 2020 US election dataset

**Table 1.** *Processing time measurements (milliseconds) across multiple iterations*

| Model | Attempt 1 | Attempt 2 | Attempt 3 | Attempt 4 | Attempt 5 | Median | Standard deviation |
|---|---|---|---|---|---|---|---|
| **RCNN** | 125 | 124 | 126 | 123 | 127 | 125 | 1.6 |
| **RCNN+NBC** | 131 | 132 | 130 | 132 | 131 | 131.2 | 0.8 |

*Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 4 (34)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

The processing time results show moderate differences between the model architectures. The baseline RCNN implementation achieved the lowest average processing time (125 ms). The RCNN+NBC configuration required 5.0% more time (131.2 ms) compared to the baseline.

To evaluate information efficiency, each architecture was evaluated using progressively larger training datasets until an precision of over 80% was achieved. This threshold was set based on expert assessment as the minimum acceptable level of performance for practical implementation. The results reveal significant differences in data efficiency between the options presented. The RCNN+NBC configuration demonstrates exceptional data efficiency, requiring only 500 samples to achieve acceptable performance – a 90% reduction compared to the baseline implementation, which required 5,000 samples.

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

**Table 2.** *Normalized experiment results*

| Model | Saving time on information processing | Precision | Information volume savings |
|---|---|---|---|
| **RCNN** | 1.00 | 0.65 | 0.00 |
| **RCNN+НБК** | 0.95 | 0.95 | 0.90 |

The application of linear additive convolution together with the weights defined above for each of the target indicators determines efficiency coefficients of 0.62 for simple RCNN and 0.945 for RCNN+NBC.

The experimental results demonstrate that the RCNN+NBC model showed an average 52.5% increase in efficiency compared to the direct application of the RCNN method. This improvement covers all evaluated metrics, with particularly significant improvements in data efficiency and classification accuracy. It follows that this model is more effective because it achieves the highest overall convolutional function value of 0.945. The proposed model combines the reliable feature extraction capabilities of RCNN with the probabilistic classification structure of the Bayesian approach, resulting in exceptional data efficiency while maintaining high classification accuracy.

## Conclusions

The aim of the study was to develop an effective two-layer model for detecting text information forgery based on a hybrid recurrent-convolutional neural network approach and a naive Bayesian classifier. The study conducted a comprehensive analysis of the characteristics of text information falsification in socially oriented systems, which are determined by significant user load. Based on expert assessment, a classification structure was formed, containing five categories of fake information – from satire to globally harmful news. In addition, a set of seven discriminatory features has been developed to identify fabricated information: emotional characteristics, rhetorical density coefficient, frequency of negative constructions, contextual emotional index, suspicion coefficient, message impact factor, and mood vector. These features form the basis for classification through a naive Bayesian classifier, which is the first layer of the proposed model.

To improve computational efficiency, the training and information processing processes were parallelized using MapReduce technology based on the Hadoop platform. This made it possible to distribute the training of CNN components across multiple nodes with subsequent aggregation of results through a reducer. Two datasets were experimentally tested: news about the Russian-Ukrainian war (20,000 records) and the 2020 US elections (equivalent volume). A multi-criteria approach with weighting coefficients was used to evaluate effectiveness: accuracy (0.8), time savings (0.1), and data efficiency (0.1).

The results of the experiments demonstrate the significant advantages of the proposed two-layer approach. The RCNN+NBC model achieved an accuracy of 95.3% compared to 65.4% for the baseline RCNN, representing a 15% increase in performance. Particularly significant is the improvement in data processing efficiency – the two-layer model requires only 500 training samples to achieve acceptable accuracy, compared to 5,000 for the baseline architecture, representing a 90% reduction in information.

Processing time increased insignificantly (5.0%), which is offset by a significant improvement in classification quality. The overall efficiency coefficient of the two-layer model was 0.945 compared to 0.62

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 4 (34)

for the baseline implementation, demonstrating a 52.5% improvement.

The use of a two-layer classification model successfully expands the capabilities of basic detection of information falsification, in particular, assessing the scale and analyzing the intentionality of fabrication. The results confirm the feasibility of implementing the proposed approach to reduce the impact of disinformation in socially oriented systems, especially during crises. Prospects for further research include extending the methodology to multimodal content (video, images), exploring the possibilities of transfer learning between different domains of disinformation, and optimizing the architecture for real-time operation in highly loaded systems.

## Acknowledgements

The author would like to thank the Armed Forces of Ukraine for providing the opportunity to write this work and conduct research during the full-scale invasion of Ukraine by the Russian Federation.

## References

1. 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 (1). DOI: 10.1007/s13278-023-01028-5
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 27.06.2025).
3. Reis, J. C. S., Correia, A., Murai, F., Veloso, A., Benevenuto, F. (2019), "Supervised Learning for Fake News Detection", *IEEE Intelligent Systems*, No. 34(2), P. 76–81. DOI: 10.1109/MIS.2019.2899143
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: 27.06.2025).
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*, P.11–112. available at: https://dspace.mit.edu/handle/1721.1/127379 (last accessed 27.06.2025).
11. Xia, T., Chen, X. A. (2020), "A Discrete Hidden Markov Model for SMS Spam Detection", *Applied Science*, Vol. 10 (14), Article 5011. DOI: 10.3390/app10145011
12. Najar, F., Zamzami, N., Bouguila, S. (2019), "Fake News Detection Using Bayesian Inference", *Information Reuse and Integration for Data Science, 30 July – 1 August 2019, Los Angeles*, P. 389–394. DOI: 10.1109/IRI.2019.00066
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

*Received (Надійшла) 29.06.2025*
*Accepted for publication (Прийнята до друку) 30.11.2025*
*Publication date (Дата публікації) 28.12.2025*

*Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 4 (34)*
*ISSN 2522-9818 (print)*
*ISSN 2524-2296 (online)*

*Відомості про авторів / About the Authors*

**Khovrat Artem** – Kharkiv National University of Radio Electronics, Graduate Student, Department of "Software Engineering", Kharkiv, Ukraine; e-mail: artem.khovrat@nure.ua; ORCID ID: https://orcid.org/0000-0002-1753-8929; Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=58128129600

**Kobziev Volodymyr** – PhD (Engineering Sciences), Kharkiv National University of Radio Electronics, Senior Researcher, Professor Department of "Software Engineering", Kharkiv, Ukraine; e-mail: volodymyr.kobziev@nure.ua; ORCID ID: https://orcid.org/0000-0002-8303-1595; Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=6507354120

**Ховрат Артем Вячеславович –** Харківський національний університет радіоелектроніки, аспірант кафедри "Програмна інженерія", Харків, Україна.

**Кобзєв Володимир Григорович –** кандидат технічних наук, Харківський національний університет радіоелектроніки, старший науковий співробітник, професор кафедри "Програмна інженерія", Харків, Україна.

# ДВОШАРОВА МОДЕЛЬ ДЛЯ ВИЯВЛЕННЯ ФАЛЬСИФІКОВАНОЇ ІНФОРМАЦІЇ З ВИКОРИСТАННЯМ НАЇВНОГО БАЄСІВСЬКОГО КЛАСИФІКАТОРА В СОЦІАЛЬНО ОРІЄНТОВАНИХ СИСТЕМАХ

**Предметом дослідження** є проблема виявлення сфабрикованої інформації в соціально орієнтованих системах, яким властиве значне користувацьке навантаження. **Мета** – розроблення двошарової моделі класифікації фейкової інформації на основі поєднання наївного баєсівського класифікатора й гібридної рекурентно-згорткової нейромережі. У статті розв'язано такі **завдання**: експертне оцінювання та доменний аналіз для визначення базових класів фейкової інформації; аналіз лінгвістичних маркерів дезінформації та розроблення векторів ознак для класифікації; створення моделей для сегрегації даних з використанням наївного баєсівського класифікатора; експериментальна перевірка запропонованої двошарової моделі та порівняння з підходом RCNN. Упроваджено такі **методи**: аналітичний (для формування набору маркерів дезінформації); індуктивний (з метою визначення цільового набору індикаторів для реалізації другого шару моделі); експертне оцінювання (для встановлення найбільш впливових факторів ефективності та вагових коефіцієнтів ознак); експериментальний і багатокритеріальний методи оцінювання (з метою визначення найбільш ефективної моделі). **Досягнуті результати**. Сформовано класифікаційну структуру для типів фейкової інформації, що містить п'ять категорій – від жартів до глобально шкідливих новин. Розроблено набір дискримінативних ознак, властивих для сфабрикованої інформації, зокрема первинні лінгвістичні маркери та вторинні стилометричні індикатори. Визначено, що підхід з використанням двошарової моделі продемонстрував у середньому 15% підвищення ефективності порівняно з прямим застосуванням гібридної рекурентно-згортокової нейромережі. **Висновки**. Застосування двошарової моделі класифікації даних успішно розширює можливості базового виявлення факту фальсифікації інформації, зокрема оцінювання масштабу й аналіз навмисності фабрикації. Емпіричний аналіз демонструє, що імплементація двошарової моделі з наївним баєсівським класифікатором досягає середнього 15% підвищення продуктивності, на відміну від простого застосування нейронної мережі. Ця різниця в продуктивності стає особливо значущою в системах з високою пропускною здатністю, де швидка ідентифікація та реагування на сфабриковану інформацію є критичними операційними параметрами. Здобутий результат дає змогу стверджувати про доцільність упровадження запропонованого підходу, а відповідно, сприяє зменшенню впливу подібної інформації в соціально орієнтованих системах, особливо під час кризових явищ.

**Ключові слова**: аналіз даних; наївний баєсівський класифікатор; нейронні мережі; паралелізація; фальсифіковані новини.

*Бібліографічні описи / Bibliographic descriptions*

Ховрат А. В., Кобзєв В. Г. Двошарова модель для виявлення фальсифікованої інформації з використанням наївного баєсівського класифікатора в соціально орієнтованих системах. *Сучасний стан наукових досліджень та технологій в промисловості*. 2025. № 4 (34). С. 112–123. DOI: https://doi.org/10.30837/2522-9818.2025.4.112

Khovrat, A., Kobziev, V. (2025), "A two-layer model to detecting falsified information using neural networks in socially oriented systems", *Innovative Technologies and Scientific Solutions for Industries*, No. 4 (34), P. 112–123. DOI: https://doi.org/10.30837/2522-9818.2025.4.112