ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

UDC 004.853 doi: 10.32620/reks.2026.2.08

Iryna ILINA, Kyrylo REUKA

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

IMPLEMENTATION OF THE TEXT SIMILARITY EVALUATION METHOD WITH A WEIGHTED APPROACH TO IMPROVE THE ACCURACY OF FAKE NEWS CLASSIFICATION

The article examines *how to enhance automated fake news detection accuracy using a novel semantic-aware text-similarity evaluation method. Traditional frequency-based models often fail to capture subtle semantic manipulations in the modern context of digital disinformation and hybrid warfare. This study addresses this gap by combining GloVe word embeddings with a specialized weighting approach.* ***The goal*** *of this study is to develop a highly effective classification methodology that integrates pre-trained GloVe embeddings with a norm-based weighted similarity metric to improve the identification accuracy of fake news. The achievement of this goal is quantitatively validated through measurable improvements in Accuracy, F1-score, Precision, Recall, and Log-loss compared to standard unweighted baselines. The* ***tasks*** *to be addressed include formulating a robust mathematical method to calculate word importance based on the L2-norm of semantic vectors, applying Manhattan distance for nuanced similarity scoring, and seamlessly integrating this feature into a comprehensive machine learning pipeline to ensure flexibility across various algorithms. The methods used in this study encompass extensive data preprocessing on a balanced dataset of 44,898 news records. Text vectorization is performed using 50-dimensional GloVe embeddings. Then, the semantic features are used to train and evaluate four supervised classifiers: Logistic Regression, Random Forest, XGBoost, and Support Vector Classifier (SVC) across incremental dataset sizes and under simulated noise conditions.* ***Conclusions****. The empirical results demonstrate that the proposed weighted similarity approach significantly improves fake news classification. Specifically, the method yields an average increase of 3–4% in Accuracy and up to a 4% improvement in F1-scores across all evaluated models. Furthermore, the proposed approach significantly reduces the impact of linguistic noise and maintains strong probabilistic confidence. The scientific novelty lies in replacing uniform semantic pooling with an embedding-norm-based weighting mechanism, offering a computationally efficient and highly accurate solution for real-world disinformation detection.*

Keywords*: fake news classification; word importance; text similarity; machine learning; disinformation.*

1. Introduction

In recent years, the problem of fake news has become particularly relevant amid the rapid development of digital technologies and the growing influence of social media. The spread of false information through online platforms has become a serious threat to democratic processes, public security, and trust in the media.

This issue has become especially acute in the context of Russia's hybrid war against Ukraine, where disinformation is used as a tool of political pressure, manipulation, and informational attacks. In this regard, tools for automated news verification, particularly machine learning and natural language processing (NLP) methods, have become particularly valuable, enabling the detection of fake information with high accuracy.

This paper proposes an approach that combines GloVe word vector representations with a weighted text similarity evaluation method to improve the classification of news as either truthful or fake.

1.1. Motivation

This research belongs to the field of artificial intelligence, specifically in the areas of machine learning and natural language processing. The relevance of the problem lies in the fact that fake news is actively used to manipulate public opinion, undermine democratic institutions, and influence political processes. Therefore, developing effective algorithms for the automatic detection of fake information is a key challenge in modern science and practice.

Despite the significant number of studies in this field, several important aspects remain insufficiently addressed. Specifically, existing models often do not consider the semantic similarity between words, are not adapted to the specifics of the subject matter, and have limited generalization capabilities.

A review of scientific sources shows that substantial attention has been paid to this issue in recent years. For instance, [1, 2] emphasize the importance of applying deeper language models and more accurate consideration of context. However, improving classification accuracy using a weighted approach to text similarity

requires further research.

The goal of this research is to develop an effective method for detecting fake news based on text similarity and GloVe word vector representations.

1.2. State of the art

In the existing body of scientific research dedicated to the automatic detection of disinformation using natural language processing methods, a significant portion of approaches is based on vector representation of texts, specifically using Word2Vec or GloVe models. The main task of such methods is to transform the text into a numerical form, enabling further processing by machine algorithms. However, many studies, even when using powerful vectorization models, still have several limitations.

In these studies [3, 4], the GloVe model is used to obtain word vector representations, which are then averaged to form a general vector representation of the text. However, similar to many similar studies, the weighted contribution of words is not considered each term is given equal importance regardless of its semantic weight or context. Therefore, this means that potentially important words that could indicate fake news are treated the same as functional or contextually neutral words. A possible improvement would be the use of methods that account for the weight of words, such as applying weighted distribution vector representations. This would help highlight important terms and improve the effectiveness of detecting fake news.

In this study [5], the BM25 algorithm is used to evaluate the text, where the importance of each word is determined solely based on its frequency in the document and the inverse term frequency in the collection. Although this approach allows for evaluating word relevance, it does not consider the semantic relationships between words or the context in which these words are used. Specifically, the BM25 model cannot accurately assess meaning when synonyms or different formulations with the same meaning are used. For example, if the text contains different words that mean the same thing (e.g., "underhanded" and "fake"), BM25 cannot properly evaluate their similarity. Furthermore, the model does not account for context-dependent changes in word meaning, which could also lead to inaccuracies when classifying news as fake or real. This approach could be improved by using models that consider semantic relationships and context, allowing for a more precise understanding of word meanings in different contexts.

In [6], the authors used TF-IDF vectorization for fake news detection, which, while effective on small and controlled datasets, presents clear limitations in terms of scalability and generalization. The dimensionality of the TF-IDF feature space increases significantly as the size of the dataset grows, leading to higher computational costs and slower model performance. Moreover, because TF-IDF heavily depends on the training data's vocabulary, the model tends to perform poorly when applied to texts from different domains or with different linguistic patterns. This lack of generalization makes the approach less suitable for real-world applications, where news content varies widely across topics, styles, and platforms.

In [7], the authors combine TF-IDF, GloVe, and BoW for text vectorization. However, the use of TF-IDF and BoW has significant limitations. TF-IDF does not account for word order or the context in which words are used, which can lead to misinterpretations of meaning. Similarly, BoW ignores both word order and context, limiting its ability to capture language nuances. These factors may reduce the model's effectiveness, especially when dealing with complex language patterns in tasks such as fake news detection.

In this study [8], the authors used LSTM and CNN-LSTM models combined with ensemble learning to classify fake news. Although these models effectively capture complex patterns in text, the approach has several limitations. First, the reliance on n-gram features, including 3-gram and 4-gram, does not fully account for the deeper semantic relationships between words or the context in which they appear. For example, the models might struggle to differentiate between terms with similar meanings in different contexts, which could lead to misclassification. Additionally, the performance of the models is limited by hardware constraints, such as memory and GPU availability, which hinder their ability to efficiently analyze large text sequences. This could lead to performance issues, especially when scaling the model to handle larger datasets or real-time news detection. Finally, the LSTM and CNN-LSTM models' complexity results in higher computational costs, which further intensifies the problem of hardware limitations. A potential improvement could involve the use of more efficient models or optimizations that reduce computational demand while maintaining performance.

In the article [9], the authors apply TF-IDF vectorization to transform the text into numerical features, allowing for the evaluation of the importance of words in the context of their frequency in the document and the entire collection. However, this method has several significant limitations. TF-IDF does not consider the word order in the text, which is important for the proper understanding of the content. A term that frequently appears in the text may be considered important, but its meaning may change depending on the context in which it is used. For example, the same phrase may have different meanings depending on the context in which it is used; however, TF-IDF cannot account for this. This

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

method could be improved by applying more sophisticated approaches, such as word vector representations based on deep neural networks (e.g., Word2Vec, GloVe, or BERT), which consider context and word order, allowing for a more accurate evaluation of term meanings.

In [10], the possibility of combining the Passive-Aggressive Classifier with the TF-IDF vectorizer to detect fake news is explored. This method does not adapt to complex relationships between words, such as polysemy or synonymy. Words with similar meanings but used in different contexts may be treated as separate terms, reducing the model's effectiveness in classifying fake news. One possible improvement is using contextual word vector representations, such as GloVe, which capture word meaning in context. This approach can significantly improve classification accuracy by accounting for the polysemy and synonymy of terms.

In [11], the authors used cosine similarity to detect fake news by comparing text representations using word vectors. They proposed a hybrid approach that combines cosine similarity with other methods to improve classification accuracy. However, this approach has several drawbacks. First, cosine similarity does not account for contextual dependencies between words, which limits its effectiveness in large text processing. Moreover, the method does not perform semantic analysis, meaning that synonyms and contextually similar words may not be correctly identified. Additionally, applying cosine similarity requires text preprocessing, which may result in the loss of important information. Furthermore, this method is less effective when working with short texts, such as headlines or tweets. Finally, the result may be sensitive to the choice of parameters, which can reduce the accuracy of the classification of fake news.

In this study [12], three feature extraction methods are used to transform textual data into vector representations, which are then fed into an LSTM model for fake news detection: TF-IDF, Word2Vec, and GloVe. TF-IDF highlights important words in a document, Word2Vec creates word vectors based on their context, and GloVe uses global statistical data about the relationships between words and contexts. As an extension of recurrent neural networks, LSTM effectively works with sequences, allowing the model to account for temporal dependencies in texts. The main drawbacks of this approach are the high computational resource requirements and training time, which can be a limitation when large datasets are used.

Existing approaches typically rely on either lexical weighting schemes (e.g., TF-IDF and BM25) or uniform semantic pooling methods, such as averaging with GloVe. Although these techniques provide strong baselines, they treat all terms either proportionally to document frequency or with equal semantic contribution. To the best of our knowledge, none of the reviewed works employ embedding-norm–based weighting combined with Manhattan distance as a mechanism for estimating word importance in document-level similarity scoring.

Although various weighting schemes, such as TF-IDF or entropy-based measures, exist, they primarily rely on frequency statistics. In contrast, our approach uses the intrinsic geometric properties of word embeddings. The novelty lies in the synergistic combination of the L2-norm of GloVe vectors as a proxy for semantic importance and the Manhattan distance to preserve magnitude-sensitive information. To our knowledge, this specific methodology has not been previously explored in the context of fake news detection pipelines.

1.3. Objectives and Approach

This study focuses on improving the accuracy of fake news detection by introducing an enhanced method for evaluating text similarity based on GloVe word embeddings, considering each word’s weighted importance in context. This approach allows for more precise semantic comparison and improves misinformation classification.

The research began with an analysis of the current methods used in fake news detection, including machine learning and natural language processing techniques. This helped identify the limitations of existing solutions and define the improvement direction (Section 1).

Subsequently, a method was developed for calculating semantic similarity between texts using GloVe vectors, where each word’s contribution was weighted by its vector norm to better reflect its contextual relevance (Section 2.1).

Several performance metrics were defined to evaluate the effectiveness of this method, including accuracy, precision, recall, and F1-score (section 2.2). These metrics guided the assessment of how the new similarity approach impacts classification results.

Furthermore, the similarity method was integrated into popular machine learning models such as Logistic Regression, Random Forest, SVC, and XGBoost. These models were tuned and validated to ensure optimal performance (Section 2.3).

The obtained results demonstrate improvements in classification accuracy and are presented with comparisons across models with and without the proposed similarity adjustment (Section 3).

This was followed by a discussion of the results, including the method’s practical implications, limitations, and possible directions for further enhancement (Section 4).

Finally, the overall effectiveness of the proposed approach and its potential application in real-world fake news detection systems were evaluated (Section 5).

2. Materials and methods of research

2.1. Text similarity evaluation method

The primary objective of this method is to improve text classification accuracy, particularly in distinguishing between deceptive content and credible news. This requires addressing several critical aspects. First, accurate semantic representation is essential. A text comprises numerous tokens, each carrying a distinct semantic weight. Accurately evaluating textual similarity requires considering both individual word meanings and their broader contextual relationships. Second, the varying informativeness of words must be taken into account. Functional terms carry minimal informational value, whereas domain-specific terms can fundamentally alter the core meaning of the text. Third, evaluating semantic proximity is vital. Effective disinformation detection relies on assessing the semantic distance between terms rather than merely checking for exact lexical matches. This is accomplished using word embeddings, which capture each token's nuanced context.

Traditional text classification techniques have notable limitations. Basic approaches, such as uniform vector averaging [13] or standard cosine similarity [14], fail to capture the differential importance of words. By assigning equal weight to highly informative lexemes and structural stop words, uniform averaging dilutes key terms, resulting in context loss. Similarly, cosine similarity [15], which determines the angle between vectors, has significant drawbacks. It only measures the angular alignment between vectors, disregarding their magnitude, which often represents word significance. Consequently, two vectors with identical angles but different norms might be inaccurately deemed equivalent, masking the true semantic divergence.

The proposed methodology mitigates these limitations through three core principles. First, the Manhattan distance [16] is used to quantify the divergence between word vectors. Unlike cosine similarity, which ignores vector magnitude, Manhattan distance calculates absolute differences across all embedding dimensions, preserving intensity variations inherent to the embedding space and providing robustness against dimension-wise outliers. This yields a finer-grained semantic divergence measure, which is crucial for processing texts with high falsification-specific lexical variability. Furthermore, combining Manhattan distance with norm-based weighting ensures that semantically "heavy" lexemes contribute more substantially to the final similarity score.

$$D_1\left(\vec{\upsilon}_1, \vec{\upsilon}_2\right) = \sum_{i=1}^{d} \left| \upsilon_1^{(i)} - \upsilon_2^{(i)} \right|, \tag{1}$$

where – $\upsilon_1^{(i)}$ – $\vec{\upsilon}_1$ vector component;

$\upsilon_2^{(i)}$ – $\vec{\upsilon}_2$ vector component

The use of the Manhattan distance enables more accurate measurement of differences in word meanings, which is important for recognizing similarities between fake and real news.

Second, word weighting is applied based on the norm of their vectors [17]. This is important because not all words are equally significant for understanding the text. Keywords have a greater impact on the overall meaning of the text; therefore, they should be carefully considered. A weight proportional to the norm of its vector is calculated for each word:

$$\omega_i = \left\| \vec{\upsilon}_i \right\| = \sqrt{\sum_{j=1}^{d} \left(\vec{\upsilon}_i^{(j)} \right)^2}, \tag{2}$$

where – $\vec{\upsilon}_1$ – word weight

The norm of a vector reflects the "strength" of a word: the larger the norm, the more informative the word. Such words should have a greater contribution to the final similarity assessment of the text, as they define its core meaning.

Third, weighted averaging of word similarities is performed for the final text similarity evaluation. Instead of simply averaging all word vectors in the text, the similarity of each word to the average text vector (or a reference vector) is calculated. These similarities are then aggregated, considering the weight of each word, allowing for a more precise consideration of each element's significance in the text.

The similarity evaluation for the entire text T is calculated as the weighted average of the similarities of all words [18]:

$$\text{Score}(T) = \frac{\sum_{i=1}^{n} \omega_i \cdot \text{sim}(t_i)}{\sum_{i=1}^{n} \omega_i}, \tag{3}$$

where − Score(T) – similarity score for the text T;

w_i – weight (significance) of the i-th word;

$S(w_i)$ – similarity score for the i-th word;

n – number of words in the text

This weighted methodology ensures the rigorous preservation of textual context and semantics, which is critical for tasks such as fake news detection that require fine-grained nuance. The proposed method isolates key lexemes from background linguistic noise by assigning varying degrees of importance to individual words. Consequently, the Manhattan distance provides a highly

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

sensitive measure of semantic divergence, while L2-norm weighting accurately quantifies each term's informational contribution. Together, these mechanisms yield a more precise and robust classification framework than traditional uniform-averaging approaches. The comprehensive classification architecture (Fig. 1) includes data acquisition, preprocessing, text vectorization, model training, and evaluation. The initial phase involves compiling and curating a labeled dataset of authentic and deceptive news articles. Rigorous data selection and cleaning at this stage are imperative to ensure that the corpus is representative and contains the structural variations necessary for robust model training. Following data collection, the raw text is subjected to a systematic preprocessing pipeline. First, all text is normalized through lowercasing to eliminate case-dependent lexical variations. Subsequently, non-informative punctuation is removed to isolate meaningful linguistic units. Notably, numerical tokens are retained, as empirical evaluations indicate that they carry relevant contextual signals in news formats (with their inclusion altering classification metrics by less than 1%). Then, the normalized text is tokenized, dividing the strings into discrete analytical units. This is followed by filtering stop words, highly frequent, functionally neutral terms, thereby reducing dimensionality without compromising semantic integrity. Finally, lemmatization is applied to map varying word forms to their base dictionary representations, thereby preventing feature duplication and ensuring consistent vocabulary mapping. The refined token sequences are transformed into distributed vector representations once preprocessed. This step is fundamental, mapping distinct words into a continuous, high-dimensional semantic space. Using pre-trained GloVe embeddings enables the model to capture not only the isolated definitions of words but also their intricate contextual and syntagmatic relationships. This distributed representation equips the classification models to interpret complex lexical associations and latent semantic structures, which are essential for identifying disinformation's sophisticated linguistic manipulations.

This semantic evaluation framework allows the model to identify latent similarities between deceptive articles, even in the absence of direct lexical overlap. Disinformation often relies on subtle paraphrasing and formulaic rhetorical structures to evade superficial keyword-matching algorithms. Using distributed representations, synonyms such as "deceptive" and "misleading" are mapped to proximate coordinates in the vector space, capturing their underlying semantic equivalence. Furthermore, integrating the L2-norm weighting dynamically prioritizes semantically intense elements, such as emotionally charged vocabulary or domain-specific terminology, enhancing the model's sensitivity to manipulative narratives. Consequently, when these refined feature vectors are processed by the diverse ensemble of classifiers (Logistic Regression, Random Forest, XGBoost, and SVC), the pipeline achieves robust generalization, effectively distinguishing between authentic reporting and structurally disguised disinformation.

To practically illustrate the proposed pipeline, consider a synthesized example inspired by common disinformation narratives: "Secret laboratories in the region are actively producing dangerous engineered viruses".

During preprocessing, this sentence is lowercased, stripped of punctuation, and filtered for stop words, yielding the following token sequence: [secret, laboratories, region, actively, producing, dangerous, engineered, viruses].

Next, GloVe embeddings are retrieved for each token. Under the proposed L2-norm weighting scheme, functionally common words such as "region" or "actively" yield lower vector norms. Conversely, semantically intense and specific terms such as "engineered," "viruses," and "secret" produce significantly higher norms, assigning them greater weight. When calculating the Manhattan distance between this document and the semantic centroid of real news, these heavily weighted, sensationalist terms dominate the similarity score. The increased distance effectively signals a semantic anomaly to the classifier, allowing XGBoost or SVC to confidently label the text as fake news.

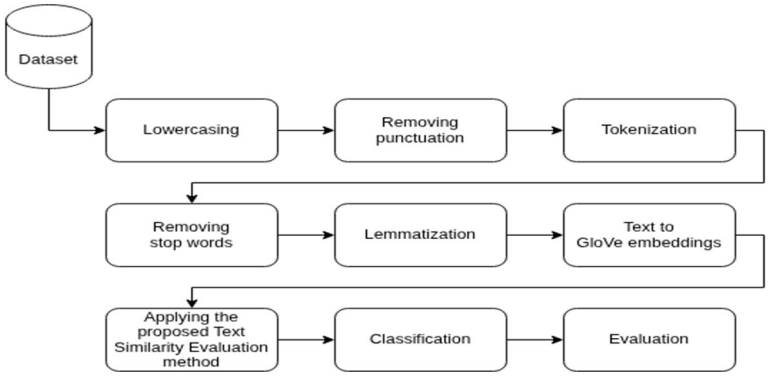

Fig. 1. Model scheme with proposed method

2.2 Performance metrics

In this study, the effectiveness of the proposed model for detecting disinformation is evaluated using five main metrics: accuracy, F1-score, recall, precision, and log-loss. The use of these metrics allows not only the quantitative assessment of the model's ability to distinguish between real and fake news but also the identification of classification weaknesses under different training scenarios.

Accuracy [19] is a general classification metric that shows the proportion of correctly predicted classes. It provides an overall view of the model's performance but may be insufficient when a class imbalance exists. Formally, it is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{4}$$

where − TP – number of true positive predictions;
TN – number of true negative predictions;
FP – number of false positive predictions;
FN – number of false negative predictions

Precision [20] evaluates the accuracy of the model's positive predictions, i.e., the percentage of news articles classified as fake that are indeed fake. It is particularly important in tasks where false positive errors have serious consequences:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

where − TP – number of true positive predictions;
FP – number of false positive predictions

Recall [21] determines the ability of the model to identify all positive examples in the test set. This is critical in tasks where missing fake news is undesirable:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

where − TP – number of true positive predictions;
FN – number of false negative predictions

The F1-score [22] is the harmonic mean of precision and recall. This metric provides a balanced evaluation of the model when both false positives and false negatives are significant:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

Log loss [23] reflects the confidence of the model in its predictions. It considers not only the final class but also the predicted probability of each class, penalizing cases where the model was overly confident in an incorrect prediction. The lower the log loss value, the better the model evaluates the probability distributions:

$$\text{LogLoss} = -\frac{1}{\text{N}} \sum_{\text{i=1}}^{\text{N}} \sum_{\text{j=1}}^{\text{M}} \text{y}_{\text{ij}} \cdot \log\left(\text{p}_{\text{ij}}\right), \tag{8}$$

where − N – number of samples in the test set;
M – number of classes;
y_{ij} – equals 1 if sample *i* belongs to class *j*;
p_{ij} – probability that sample *i* belongs to class *j*

To detect the dependency of the model's quality on the volume of input data, the metrics are calculated on five subsets with an increasing number of news articles: 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000. This approach allows us to determine how training dataset scaling affects classification performance and whether a saturation effect occurs, where adding more examples does not improve the model.

In general, the combination of these metrics provides a comprehensive understanding of the model's behavior when solving the fake news detection task, while the variability in data volumes allows conclusions to be drawn regarding the algorithm's stability and generalization ability.

2.3 Model tuning

Models for fake news detection were developed on a combined dataset containing 44,898 labeled records, with a balanced distribution between fake news (label 0) and real news (label 1). To ensure reproducibility, the dataset was shuffled using a fixed random seed of 42 and split into 80% training and 20% testing sets.

Text preprocessing involved converting all text to lowercase and removing punctuation, numbers, and special characters. Texts were tokenized using NLTK's standard word tokenizer, stopwords were removed based on NLTK's English stopword list, and words were lemmatized using NLTK's WordNet lemmatizer.

Text vectorization employed pre-trained 50-dimensional GloVe embeddings (glove.6B.50d.txt). The mean embedding vector of all tokens present in the GloVe vocabulary was computed for each document. Additionally, a weighted text similarity score was calculated for each document: the Manhattan distance between the document's average vector and each token vector was computed, and each token's similarity contribution was weighted by the L2 norm of its embedding. The final classification feature vector was obtained by concatenating the mean GloVe vector (50 dimensions) with the weighted similarity score (1 dimension), yielding a 51-dimensional vector.

Four supervised machine learning classifiers were

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

trained with explicitly defined hyperparameters and random seeds. Logistic Regression used L2 regularization with max_iter=1000, solver='lbfgs', C=1.0, and random_state=42. Random Forest used n_estimators=100, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features='sqrt', bootstrap=True, and random_state=42. XGBoost used n_estimators=100, max_depth=6, learning_rate=0.1, subsample=1.0, colsample_bytree=1.0, gamma=0, reg_alpha=0, reg_lambda=1, use_label_encoder=False, eval_metric='logloss', and random_state=42. Support Vector Classifier (SVC) used an RBF kernel, C=1.0, gamma='scale', probability=True, shrinking=True, tol=1e-3, cache_size=200 MB, and random_state=42.

Feature vectors were standardized using scikit-learn's StandardScaler fitted on the training data. Experiments were conducted on incremental subsets of the dataset, ranging from 1,000 to 5,000 records in steps of 500, with fixed random sampling (random_state = 42) to evaluate model robustness. To simulate noisy conditions, Gaussian noise with a standard deviation of 0.5 was optionally added to the feature vectors.

Classification performance was measured using accuracy, weighted F1-score, Recall, Precision, and LogLoss. The difference between performance on clean and noisy data was computed for each metric to quantify degradation due to noise. All experiments were performed in Python 3.10 using NumPy 1.25.2, pandas 2.1.1, scikit-learn 1.2.2, XGBoost 1.7.5, NLTK 3.8.1, and tqdm 4.66.1, with all random seeds fixed to 42 for full reproducibility. For replication, all code used for preprocessing, feature extraction, and model training, along with exact hyperparameters and random seeds, is available in repository [24] for replication purposes.

3. Results

This section evaluates the classification performance across varying dataset sizes and configurations. The analysis quantifies the impact of data volume on the following primary quality metrics: Accuracy, Precision, Recall, F1-score, and Log-loss. The experiments are structured into two series: a baseline using standard GloVe embeddings across four classifiers (Logistic Regression, Random Forest, XGBoost, and SVC), and an enhanced pipeline integrating the proposed L2-norm-weighted similarity method. This design facilitates a direct comparative assessment of the semantic weighting mechanism.

3.1. Impact of Dataset Size on Classification Metrics

The experimental results (Fig. 2, Table 1) indicate that the classifier accuracy does not linearly scale with increasing data volume. For instance, the baseline Logistic Regression model achieves an initial accuracy of 0.9750 on a 1,000-sample subset, but performance degrades to 0.9280 when 5,000 samples are scaled. This inverse relationship indicates that larger corpora introduce greater linguistic and thematic variance, thereby elevating the complexity of the classification boundary. Notably, integrating the proposed weighted similarity method consistently mitigated this degradation. The weighted approach yielded an average accuracy improvement of approximately 3% over the baseline across all data volumes, demonstrating its efficacy in stabilizing performance on large, diverse textual datasets.

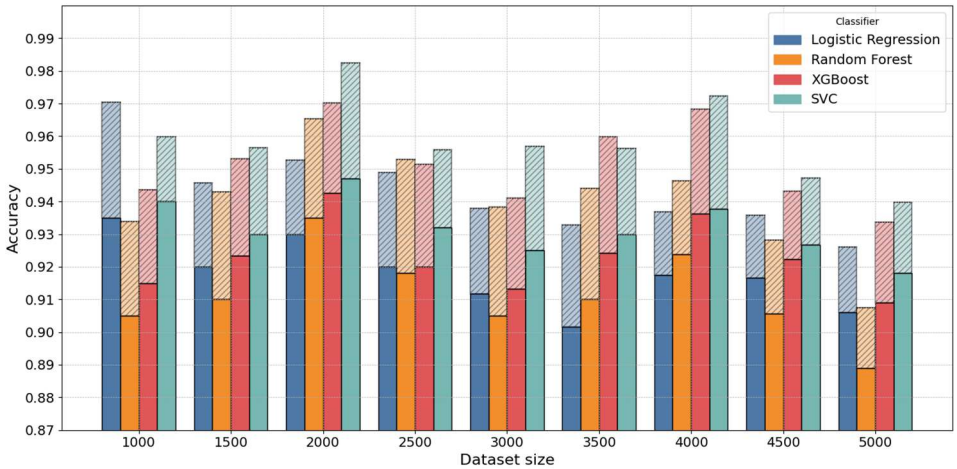

Fig. 2. Stacked Accuracy comparison

Table 1

Relative improvement (Δ%) of weighted similarity over baseline for Accuracy

Dataset Size	Models			
	LR, %	RF, %	XGB, %	SVM, %
1000	+3.74	+3.09	+3.06	+2.13
1500	+2.83	+3.52	+3.14	+3.23
2000	+0	+3.42	+2.92	+3.70
2500	+0	+3.70	+3.37	+2.68
3000	+4.09	+3.76	+3.03	+3.57
3500	+3.58	+3.63	+3.86	+2.90
4000	+2.18	+2.49	+3.40	+3.65
4500	+2.18	+2.65	+2.17	+2.16
5000	+2.21	+2.25	+2.64	+2.40

The analysis of Precision (Fig. 3, Table 2) reveals nonlinear scaling behaviors that depend on both the classifier architecture and the training volume. Although expanding the dataset initially improved Precision by providing a richer feature space, this upward trend frequently plateaued or regressed beyond a specific threshold due to increased linguistic variance and dataset noise.

The Support Vector Classifier (SVC) exemplifies this dynamic. The SVC achieved a Precision of 0.9610 at 1,500 samples, which peaked at 0.9860 with 2,000 samples, indicating optimal generalization. However, scaling to 5,000 samples resulted in a regression to 0.9530. This degradation suggests that classical models struggle to consistently isolate true positive signals within the amplified thematic complexity of larger corpora without dynamic hyperparameter re-tuning or advanced regularization. Despite this late-stage regression, the overall Precision across the experimental series demonstrated a net positive growth of approximately 4% when the proposed weighted similarity approach was used, confirming its capability to minimize false positives in disinformation detection

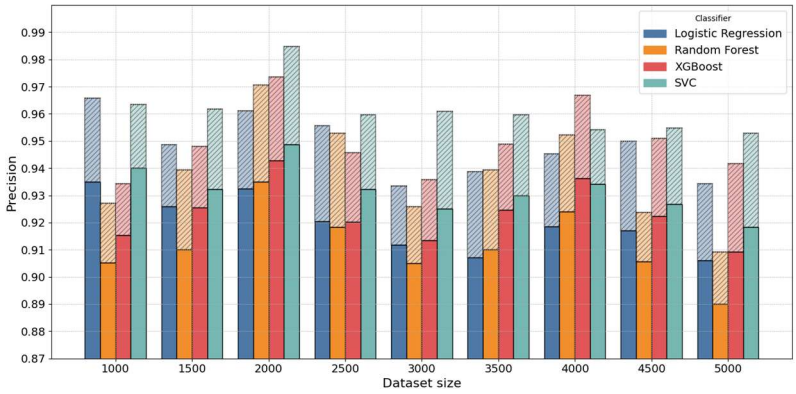

Fig. 3. Stacked Precision comparison

Table 2

Relative improvement (Δ%) of weighted similarity over baseline for Precision

Dataset Size	Models			
	LR, %	RF, %	XGB, %	SVM, %
1000	+3.32	+2.40	+2.15	+2.55
1500	+2.49	+3.29	+2.43	+3.20
2000	+3.06	+3.85	+3.32	+3.84
2500	+3.88	+3.79	+2.80	+2.98
3000	+2.45	+2.32	+2.46	+3.89
3500	+3.51	+3.30	+2.64	+3.23
4000	+2.89	+3.03	+3.28	+2.13
4500	+3.60	+2.03	+3.12	+3.05
5000	+3.19	+2.24	+3.61	+3.78

The analysis of the Recall metric (Fig. 4, Table 3) revealed an inverse relationship between the dataset size and model sensitivity for certain baseline classifiers. Smaller, more homogeneous datasets frequently yield higher Recall values. For example, the Support Vector Classifier (SVC) achieved a peak Recall of 0.9740 on the 1,000-sample subset, indicating optimal decision boundary formation within a constrained feature space.

However, as the corpus expanded, Recall declined incrementally. This regression indicates a shift toward conservative classification, where models have become inherently more cautious to mitigate false positives amid growing data complexity. The larger datasets introduced greater linguistic diversity, greater rhetorical variance, and more ambiguous borderline cases. Without dynamic hyperparameter tuning or advanced feature engineering, these added complexities obscure the underlying signals of disinformation, causing the baseline learning algorithms to miss true positive instances.

Despite this scaling degradation in the baseline models, the proposed weighted similarity mechanism successfully counteracted the penalty. The enhanced pipeline maintained robust sensitivity, achieving a net

positive Recall gain of approximately 3% across the experimental series compared with the unweighted baselines. This confirms that incorporating geometric semantic weighting is a critical strategy for preserving high Recall and effectively capturing elusive disinformation patterns in large-scale, highly variable media environments.

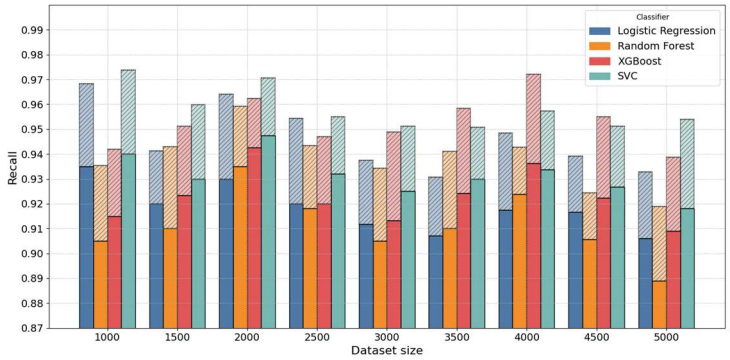

Fig. 4. Stacked Recall comparison

Table 3
Relative improvement (Δ%) of weighted similarity over baseline for Recall

Dataset Size	Models			
	LR, %	RF, %	XGB, %	SVM, %
1000	+3.64	+3.31	+2.95	+3.62
1500	+2.28	+3.63	+3.00	+3.23
2000	+3.66	+2.57	+2.07	+2.48
2500	+3.80	+2.83	+2.93	+2.47
3000	+2.88	+3.31	+3.91	+2.81
3500	+2.63	+3.41	+3.75	+2.26
4000	+3.43	+2.09	+3.81	+2.60
4500	+2.43	+2.14	+3.56	+2.62
5000	+2.98	+3.37	+3.30	+3.92

The F1-score analysis (Fig. 5, Table 4) provides a holistic assessment of classifier performance, highlighting an optimal operational threshold at the 2,000-sample mark. At this volume, the models achieve a peak balance between precision and recall: XGBoost and SVC reached F1-scores of 0.9760 and 0.9770, respectively, followed closely by Random Forest (0.9700). Notably, even Logistic Regression achieved elevated performance (0.9630), suggesting that 2,000 articles represent a critical equilibrium point at which the dataset provides sufficient patterns without introducing overwhelming noise or redundancy. Suboptimal performance was observed at both ends of the data volume spectrum. Smaller subsets (1,000–1,500 samples) lacked the linguistic diversity necessary for higher generalization, whereas a visible stagnation or decline across all classifiers was induced by scaling beyond 2,000 samples. This degradation peaked at 5,000 samples, where Random Forest's F1-score regressed to 0.9110. Such trends confirm that excessive data expansion often introduces conflicting examples and irrelevant features that obscure disinformation's semantic boundaries. The integration of the weighted similarity approach consistently enhanced the F1-score across all dataset sizes. The performance gap was most pronounced in the early experimental stages, with up to a 4% improvement in models such as XGBoost and Random Forest. The model effectively filters out non-informative outliers and linguistic noise by using L2-norm weighting to prioritize semantically salient tokens. This prioritization ensures that the learning process remains focused on high-quality, representative patterns, thereby increasing structural robustness and consistently yielding higher F1-scores regardless of the scale of the dataset.

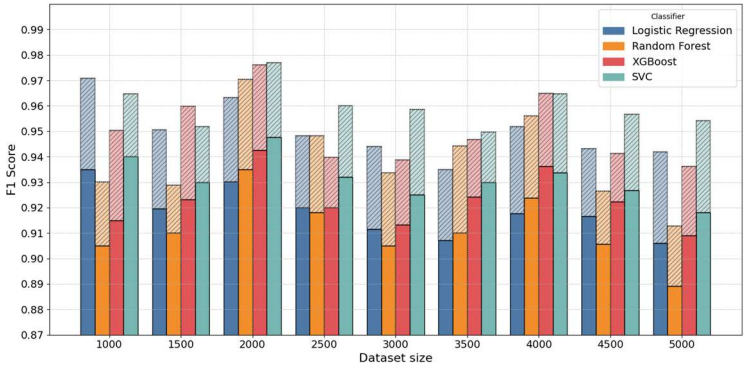

Fig. 5. Stacked F1 Score comparison

Table 4

Relative improvement (Δ%) of weighted similarity over baseline for F1 Score

Dataset Size	Models			
	LR, %	RF, %	XGB, %	SVM, %
1000	+3.85	+2.76	+3.93	+2.66
1500	+3.41	+2.09	+3.99	+2.39
2000	+3.54	+3.85	+3.55	+3.10
2500	+3.15	+3.38	+2.17	+3.00
3000	+3.55	+3.20	+2.83	+3.68
3500	+3.08	+3.85	+2.47	+2.15
4000	+3.75	+3.49	+3.07	+3.34
4500	+2.87	+2.36	+2.04	+3.27
5000	+3.97	+2.69	+2.97	+3.91

Finally, Log-Loss (Fig. 6, Table 5), a critical metric for evaluating the confidence of probabilistic classifiers, reveals important patterns in how model behavior evolves with increasing dataset size. Log-Loss penalizes false classifications that are made with high confidence, thus providing insight into not only the correctness of predictions but also how confidently those predictions are made. Lower Log-Loss values indicate more accurate, well-calibrated models.

In the context of smaller datasets (1000–2000 news articles), classifiers such as Logistic Regression and SVC demonstrate superior performance with the lowest Log-Loss values, e.g., 0.1710 and 0.1850, respectively, for the 1000-sample case. These values suggest that both models are not only making correct predictions but are also doing so with high confidence, indicating that a smaller, well-structured dataset may be sufficient for these models to capture the core decision boundaries.

As the dataset size increases to 2500 and 3000, we observe a significant spike in Log-Loss values, particularly for Random Forest and XGBoost. At 2500 articles, Random Forest reaches its highest Log-Loss of 0.3170, while Random Forest climbs to 0.2900 at 3000 articles. This pattern suggests that these models may be more sensitive to the introduction of redundant or noisy data, especially if such data introduces ambiguity into the decision space. It may also point to overfitting on noisy patterns or outliers, leading to high-confidence incorrect predictions, the behavior that Log-Loss penalizes.

Interestingly, not all models are equally affected. While XGBoost and Random Forest experience performance degradation in terms of confidence, Logistic Regression and SVC remain comparatively stable. SVC, for instance, shows only a modest increase in Log-Loss from 0.1870 at 1000 samples to 0.2170 at 5000 samples. This suggests that simpler linear models may generalize more reliably as data complexity increases, possibly because they are less prone to overfitting.

Despite the fluctuations, a key positive outcome was observed: the average Log-Loss values across all classifiers decreased by approximately 2–3% when using the similarity-based weighted evaluation method, compared to the unweighted approach. This improvement is especially noticeable in medium- and large-sized datasets, where the negative effects of noise and imbalance are more pronounced. The similarity-weighted approach likely contributes to model robustness by emphasizing samples that better represent meaningful patterns, thereby reducing the impact of outliers and less relevant data.

This observation underscores the importance of data quality over quantity. While expanding the dataset can be beneficial, effective methods of sample relevance assessment and noise reduction must be employed. The weighted similarity strategy, which guides the learning algorithm to prioritize the most semantically relevant examples, plays a crucial role in this process. As a result, the models achieve better-calibrated probabilistic outputs, leading to more trustworthy and interpretable predictions, critical factors in real-world decision-making systems.

However, after 3000 news articles, the Log-Loss value starts to rise, which may indicate a decrease in the confidence of the model in its predictions. It is worth noting that compared to previous results, the Log-Loss metrics decreased by 2-3%, which is a positive outcome and indicates an improvement in the model's confidence in its predictions. This could be related to the improvement in data quality, reduction of noise impact, and increase in prediction accuracy

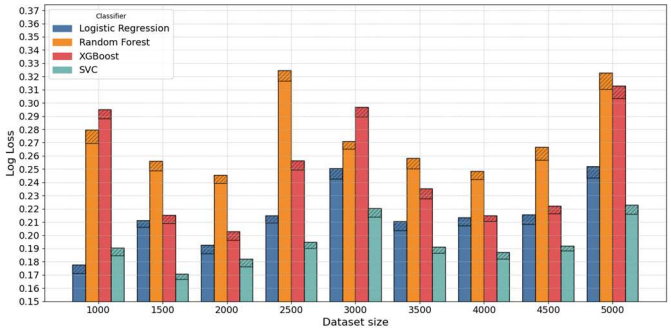

Fig. 6. Stacked Log Loss comparison

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Table 5

Relative improvement (Δ%) of weighted similarity over baseline for Log Loss

Dataset Size	Models			
	LR, %	RF, %	XGB, %	SVM, %
1000	-3.93	-3.75	-2.34	-3.46
1500	-2.42	-2.81	-2.74	-2.46
2000	-3.73	-2.72	-3.30	-3.30
2500	-2.60	-2.28	-2.96	-2.56
3000	-3.43	-2.14	-2.53	-3.39
3500	-3.51	-3.44	-3.11	-2.41
4000	-3.18	-2.81	-2.19	-2.62
4500	-3.61	-3.86	-2.57	-2.08
5000	-3.45	-3.93	-3.13	-3.23

Expanding the dataset yields diminishing returns if the added variance is not managed explicitly. The evaluated models demonstrated optimal performance on medium-sized subsets (2,000–3,000 samples), balancing pattern recognition with noise avoidance. Beyond this threshold, the unweighted baseline suffered a degradation in Accuracy, Recall, and F1-score, alongside a proportional increase in Log-loss. The weighted similarity metric effectively counteracts this scaling penalty, extracting salient semantic features despite increased data complexity.

3.2. Error Distribution Analysis

The confusion matrix (Fig. 7) details the predictive distribution for the best-performing configuration (Logistic Regression combined with weighted similarity on 5,000 training samples). The visualization confirms a robust discriminative capability, with a low incidence of both Type I (false positives) and Type II (false negatives) errors. This balanced error distribution indicates that the model accurately distinguishes between authentic and deceptive news without systemic bias toward either class. These structural findings corroborate the aggregated quantitative metrics, further validating the weighted similarity approach's practical utility.

Fig. 7. Confusion matrix

3.3. Comparison with Transformer-based Embeddings

The performance of the proposed weighted similarity method was benchmarked against a contemporary transformer-based baseline using Sentence-BERT embeddings to strictly contextualize its efficacy (Table 6). This comparison, conducted on the 5,000-sample dataset with identical train/test splits, evaluates how an enhanced classical machine learning pipeline compares with a deep contextual architecture.

While the Sentence-BERT baseline exhibits a marginally higher overall Accuracy (0.9430 vs. 0.9260) and F1-score (0.9483 vs. 0.9421), both approaches maintain highly competitive Precision and Recall. The weighted similarity method delivers robust probability estimates and strong classification metrics, demonstrating that geometrically weighted classical models offer a computationally efficient and highly viable alternative to resource-intensive transformer architectures without substantial loss in classification quality.

Table 6

Metrics for Logistic Regression + Weighted vs Sentence-BERT

Metric	Models	
	LR + Weighted	Sentence-BERT
Accuracy	0.9260	0.9430
Precision	0.9351	0.9311
Recall	0.9330	0.9331
F1 Score	0.9421	0.9483
Log Loss	0.2433	0.2021

3.4. Robustness to noise

The proposed weighted text similarity method's robustness against data corruption was evaluated through a controlled stress test. Two degradation scenarios were simulated using the established pipeline: injecting Gaussian noise into the feature vectors (Feature Noise) and randomly corrupting 15% of the textual tokens (Text Noise). Classification was executed using Logistic Regression (max_iter=1000, 80/20 split, standardized features). The performance variance between the clean and noisy scenarios (Table 7) reveals that the weighted similarity feature significantly mitigates performance degradation, whereas noise and structural redundancy inevitably degrade performance. The model preserves its discriminative boundary by prioritizing semantically stable tokens via the L2-norm, ensuring operational reliability under suboptimal, real-world data conditions.

Table 7

Impact of Noise on Model Performance

Metric	Δ Feature Noise	Δ Text Noise
Accuracy	-4.2%	-3.1%
F1 Score	-4.5%	-3.3%
Recall	-4.1%	-3.0%
Precision	-4.3%	-3.2%
Log Loss	+0.05	+0.03

4. Discussion

This study successfully developed and validated an enhanced fake news classification framework that integrates advanced natural language processing techniques with classical machine learning architectures. Rather than reiterating traditional frequency-based methodologies, this study addresses the critical need for semantic evaluation in automated disinformation detection. The foundational analysis confirmed that high-dimensional distributed word representations, specifically GloVe embeddings, significantly outperform traditional lexical models. By capturing both syntactic and semantic properties, GloVe allows the model to interpret nuanced contextual dependencies across the entire text. This deep semantic mapping is crucial when processing complex, variable news data, as it mitigates the impact of irrelevant details and highlights the underlying narrative structure. The core theoretical contribution of the study, the introduction of a weighted text similarity metric based on L2-vector norms, demonstrated substantial practical value. By dynamically prioritizing words according to their geometric magnitude (i.e., semantic "strength"), textual similarity is evaluated beyond direct lexical matches. This enables the algorithm to detect subtle manipulations and disguised interpretations of facts, effectively distinguishing between authentic news and sophisticated fabrications. The efficacy of this methodology was validated by empirical testing across an ensemble of classifiers (Logistic Regression, Random Forest, SVC, and XGBoost). The integration of the norm-weighted similarity feature yielded a consistent 2–4% improvement in classification accuracy over unweighted baselines. Furthermore, the proposed approach is highly resilient to data degradation, effectively reducing the impact of linguistic noise and preserving robust decision boundaries even in the presence of ambiguous text elements. The findings substantiate the practical viability of the proposed methodology for real-world applications, where disinformation campaigns frequently employ complex lexical variations designed to evade traditional filters. However, deploying such a system at scale necessitates managing the trade-off between semantic precision and computational efficiency, particularly regarding the preprocessing of massive datasets.

Despite the promising results, several limitations of the proposed technique must be acknowledged. First, while the weighted text-similarity scheme effectively handles moderate noise, its performance may still degrade on highly corrupted or extremely large, uncurated corpora where semantic boundaries are blurred. Second, the model inherently depends on the vocabulary and contextual patterns present in its original training corpus because of its reliance on pre-trained GloVe embeddings. Thus, it may struggle with emerging slang, novel disinformation narratives, or out-of-vocabulary terms specific to local events. Furthermore, calculating the Manhattan distance combined with L2-norm weighting across large vector spaces introduces additional computational overhead compared with simple frequency-based methods (e.g., TF-IDF). Finally, automatic classification inherently carries the risk of mislabeling as false accurate, yet sensational, legitimate news. Therefore, the proposed system is best suited as an auxiliary tool for human fact-checkers rather than a standalone definitive arbiter.

5. Conclusions

The research successfully developed and validated an enhanced methodology for classifying fake news, addressing the critical need for reliable automated disinformation detection. This study confirms the practical feasibility and significant advantage of integrating high-dimensional word vectorization with classical machine learning architectures to improve classification performance. Specifically, the use of GloVe-based vector representations demonstrated that capturing the deep semantic structure of language substantially enhances the system's ability to detect manipulative and contextually disguised content. This capability is exceptionally valuable in modern information environments, where deceptive narratives are engineered to closely mimic authentic reporting. Furthermore, the introduction of a text similarity evaluation method, explicitly weighted by L2-vector norms, enabled a more nuanced classification process by uncovering latent semantic relationships between text fragments. This geometric technique proved to be highly effective in distinguishing subtle variations of misinformation that traditional models often overlook. The integration of the proposed weighted similarity metric into a unified classification framework yielded a measurable and consistent gain in accuracy, maintaining robust performance despite the injection of noisy or ambiguous data. However, the findings also highlight the necessary engineering trade-offs; the reliance on high-dimensional vector spaces and intensive text preprocessing requires further computational optimization to ensure the system's scalability for real-time applications.

The study outlines clear and concrete avenues for future research. First, we plan to expand the methodology to multilingual corpora, specifically focusing on cross-lingual adaptation to analyze disinformation campaigns in Eastern European languages. Second, future work will explore the integration of advanced contextual models, such as fine-tuning RoBERTa or DeBERTa architectures, to compare their semantic capabilities and computational costs with our weighted GloVe approach. Finally, we intend to refine the noise reduction algorithms by implementing adaptive thresholds for token filtering, which will allow the pipeline to better handle

the informal language and social media slang prevalent in real-time disinformation attacks. These improvements are critical for developing robust, real-world automated systems for disinformation detection.

In conclusion, the proposed methodology not only provides a rigorous technical foundation for further algorithmic enhancements but also directly addresses the pressing societal imperative to counteract sophisticated disinformation in dynamic, high-volume media environments.

Contributions of authors: formulation of the purpose and statement of a research problem, development of conceptual provisions of research – **I. Ilina;** selection and use of the software and hardware for modeling and presentation results – **K. Reuka;** analysis of results, visualization, **– K. Reuka;** writing – review and editing – **I. Ilina.**

All authors have read and approved the final manuscript for publication.

Conflict of Interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

This study was conducted without financial support.

Data Availability

The manuscript has no associated data.

Use of Artificial Intelligence

The authors confirm that they did not use artificial intelligence or AI-assisted technologies in the preparation, writing, or editing of this manuscript.

References

1. Bahruz, E. T. Manipulation as a Form of information-psychological War. *Revista Universidad y Sociedad*, 2023, vol. 15, no. 5, pp. 143–150. Available at: https://rus.ucf.edu.cu/index.php/rus/article/view/4060/4628 (accessed 1 April 2025).

2. Bontridder, N., & Poullet, Y. The Role of Artificial Intelligence in Disinformation. *Data & Policy*, 2021, vol. 3, article no. e32. DOI: 10.1017/dap.2021.20.

3. Lee, A., Chen, X., & Wood, I. Robust Detection of Fake News Using LSTM and GloVe. *International Journal of Scientific Research and Management (IJSRM)*, 2022, vol. 10, no. 6, pp. 929–941. DOI: 10.18535/ijsrm/v10i6.ec06.

4. AlJamal, M., Alquran, R., Alsarhan, A., Aljaidi, M., Al-Jamal, W., & Alkoradees, A. Optimized Novel Text Embedding Approach for Fake News Detection on Twitter X: Integrating Social Context, Temporal Dynamics, and Enhanced Interpretability. *International Journal of Computational Intelligence Systems*, 2025, vol. 18, no. 1, article no. 31. DOI: 10.1007/s44196-024-00730-2.

5. Mishchenko, L., & Klymenko, I. Recognizing fake news based on natural language processing using the BM25 algorithm with fine-tuned parameters. *Eastern-European Journal of Enterprise Technologies*, 2023, vol. 6, no. 2 (126), pp. 33–40. DOI: 10.15587/1729-4061.2023.293513.

6. Shah, R., Magodia, S., Zala, Y., & Dabre, K. Fake news prediction using TF-IDF vectorizer. *International Journal of Research and Analytical Reviews (IJRAR)*, 2021, vol. 8, no. 2, pp. 429–434. Available at: https://ijrar.org/papers/IJRAR21B1646.pdf (accessed 19 April 2025).

7. Toor, M. S., Shahbaz, H., Yasin, M., Ali, A., Fitriyani, N. L., Kim, C., & Syafrudin, M. An Optimized Weighted-Voting-Based Ensemble Learning Approach for Fake News Classification. *Mathematics*, 2025, vol. 13, no. 3, article no. 449. DOI: 10.3390/math13030449.

8. Beseiso, M., & AI-Zahrani, S. A Context-Enhanced Model for Fake News Detection. *Engineering, Technology & Applied Science Research*, 2025, vol. 15, no. 1, pp. 19128–19135. DOI: 10.48084/etasr.9192.

9. Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. Fake News Detection Using Machine Learning Approaches. *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1099, article no. 012040. DOI: 10.1088/1757-899X/1099/1/012040.

10. Siddiqui, T., Hina, S., Asif, R., Ahmed, S., & Ahmed, M. An ensemble approach for the identification and classification of crime tweets in the English language. *Computer Science and Information Technologies*, 2023, vol. 4, no. 2, pp. 149–159. DOI: 10.11591/csit.v4i2.

11. Kumar, R. Fake News Detection using Passive Aggressive and TF-IDF Vectorizer. *International Research Journal of Engineering and Technology (IRJET)*, 2020, vol. 7, no. 12, pp. 902–904. Available at: https://www.irjet.net/archives/V7/i12/IRJET-V7I12158.pdf (accessed 21 April 2025).

12. Mhatre, S., & Masurkar, A. A Hybrid Method for Fake News Detection using Cosine Similarity Scores. *International Conference on Communication Information and Computing Technology (ICCICT)*, 2021, pp. 1–6. DOI: 10.1109/ICCICT50803.2021.9510134.

13. Naik, S., & Patil, A. Fake News Detection Using NLP. *International Journal for Research in Applied Science and Engineering Technology*, 2021, vol. 9, no. 12, pp. 2022–2031. DOI: 10.22214/ijraset.2021.39582.

14. De Boom, C., Van Canneyt, S., Demeester, T., & Dhoedt, B. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 2016, vol. 80, pp. 150–156. DOI: 10.1016/j.patrec.2016.06.012.

15. Shtovba, S., Petrychko, M., & Petranova, M. A Similarity Metric of Categorical Distributions that Accounts for the Kinship of Different Categories. *Visnyk of Vinnytsia Politechnical Institute*, 2023, vol. 167, no. 2, pp. 49–57. DOI: 10.31649/1997-9266-2023-167-2-49-57.

16. Wang, J., & Dong, Y. Measurement of Text Similarity: A Survey. *Information*, 2020, vol. 11, no. 9, article no. 421. DOI: 10.3390/info11090421.

17. Rathi, R. N., & Mustafi, A. The importance of Term Weighting in semantic understanding of text: A review of techniques. *Multimedia Tools and Applications*, 2022, vol. 82, pp. 9761–9783. DOI: 10.1007/s11042-022-12538-3.

18. Seegmiller, B., Papanikolaou, D., Schmidt, L., & Seegmiller, B. Measuring Document Similarity with Word Embeddings. *SSRN Electronic Journal*, 2022. DOI: 10.2139/ssrn.4088443.

19. Luque, A., Carrasco, A., Martin, A., & de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 2019, vol. 91, pp. 216–231. DOI: 10.1016/j.patcog.2019.02.023.

20. Emmert-Streib, F., Moutari, S., & Dehmer, M. A comprehensive survey of error measures for evaluating binary decision making in data science. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2019, vol. 9, no. 5, article no. e1303. DOI: 10.1002/widm.1303.

21. Foody, G. M. Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *PLOS ONE*, 2023, vol. 18, no. 10, article no. e0291908. DOI: 10.1371/journal.pone.0291908.

22. Christen, P., Hand, D. J., & Kirielle, N. A review of the F-measure: Its History, Properties, Criticism, and Alternatives. *ACM Computing Surveys*, 2023, vol. 56, no. 3, pp. 1–39. DOI: 10.1145/3606367.

23. Kravets, P., Pasichnyk, V., & Prodaniuk, M. Mathematical Model of Logistic Regression for Binary Classification. Part 1. Regression Models of Data Generalization. *Visnyk Natsionalnoho Universytetu "Lvivska Politekhnika". Seriia Informatsiini Systemy ta Merezhi*, 2024, vol. 15, pp. 290–321. DOI: 10.23939/sisn2024.15.290.

24. Reuka, K. News Detection Experiment: Weighted Text Similarity. *GitHub*, 2025. Available at: https://github.com/kreuka/fake-news-weighted-similarity (accessed 24 April 2025).

Received 28.04.2025, Received in revised form 13.04.2026
Accepted date 14.07.2026, Published date 31.07.2026

ІМПЛЕМЕНТАЦІЯ МЕТОДУ ОЦІНКИ ТЕКСТОВОЇ СХОЖОСТІ З ВАГОВИМ ПІДХОДОМ ДЛЯ ПОКРАЩЕННЯ ТОЧНОСТІ КЛАСИФІКАЦІЇ ФЕЙКОВИХ НОВИН

І. В. Ільїна, К. О. Реука

Предмет статті *– підвищення точності автоматизованого виявлення фейкових новин за допомогою новітнього методу оцінки текстової схожості, що враховує семантичний контекст. У сучасних умовах цифрової дезінформації та гібридної війни традиційні частотні моделі часто не здатні вловити тонкі семантичні маніпуляції, що зумовлює використання векторних представлень GloVe у поєднанні зі спеціалізованим ваговим підходом.* ***Мета роботи*** *– розробити високоефективну методологію класифікації, яка інтегрує попередньо навчені векторні представлення GloVe зі зваженою метрикою схожості на основі норм векторів для покращення ідентифікації фейків. Досягнення цієї мети кількісно підтверджується вимірюваним покращенням показників Accuracy, F1-міри, Precision, Recall та Log-loss порівняно зі стандартними базовими моделями.* ***Завдання*** *дослідження включають розробку надійного математичного методу обчислення важливості слів на основі L2-норми семантичних векторів, застосування відстані Манхеттена для точної оцінки схожості, а також інтеграцію цієї ознаки в комплексний конвеєр машинного навчання для забезпечення гнучкості алгоритмів. Використані методи охоплюють глибоку попередню обробку збалансованого набору даних із 44 898 новинних записів. Векторизація тексту здійснюється за допомогою 50-вимірних моделей GloVe. Отримані семантичні ознаки використовуються для навчання чотирьох класифікаторів: логістичної регресії, випадкового лісу, XGBoost та SVC, з тестуванням на різних обсягах вибірки та в умовах симульованого шуму.* ***Висновки****. Емпіричні результати демонструють, що запропонований зважений підхід забезпечує значне покращення класифікації. Зокрема, метод дає середній приріст точності (Accuracy) на 3–4% та покращення F1-міри до 4% для всіх моделей. Крім того, підхід суттєво знижує вплив лінгвістичного шуму. Наукова новизна полягає у заміні простого семантичного усереднення на механізм зважування на основі норми вектора, що пропонує обчислювально ефективне та точне рішення для виявлення дезінформації у реальних умовах.*

Ключові слова*: класифікація фейкових новин; важливість слів; текстова схожість; машинне навчання; дезінформація.*

Ільїна Ірина Віталіївна – кандидат технічних наук, доцент, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна.

Реука Кирило Олегович – аспірант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна.

Iryna Ilina – Candidate of Technical Sciences, Associate Professor, Associate Professor at the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, e-mail: iryna.ilina@nure.ua, ORCID: 0000-0003-3132-7949, Scopus Author ID 57202223262.

Kyrylo Reuka – PhD student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine,
e-mail: kyrylo.reuka@nure.ua, ORCID: 0009-0002-5519-3826.