Neurocomputing 654 (2025) 131327

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

In depth analysis for securing the truth: Addressing the fake news challenge with graph neural networks☆

Gracjan Kątek[a], Rafał Kozik[a], Aleksandra Pawlicka[b,c], Marek Pawlicki[a], Michał Choraś[a,*]

[a] Bydgoszcz University of Science and Technology, Bydgoszcz, Poland
[b] University of Warsaw, Warsaw, Poland
[c] ITTI Sp. z o.o., Poznań, Poland

HIGHLIGHTS

- Innovative approach with a multi-factor assessment of the content of documents.
- NLP analysis to detect fake news.
- Innovative graph neural networks, which allow for a more complex and contextual understanding of the data.
- Results indicate a significant improvement in effectiveness compared to the baseline approaches.

ARTICLE INFO

Keywords:
Fake news
Machine learning
AI
Security

ABSTRACT

The fake news phenomenon has a significant impact on societies, homeland security, democracy and the functioning of the public space. The spread of false information is becoming an increasing challenge in the context of the dynamic growth of the volume of content shared by news outlets and social media. The overwhelming amount of this information makes manual verification of every news item or press release practically impossible.

The current development of technology in the field of natural language processing (NLP) opens up new possibilities for the development of automatic content verification systems. The automation of this process not only improves but also significantly speeds up the detection of unreliable information, which is a key tool in the fight against fake news.

In this article, we propose an innovative approach that involves a multi-factor assessment of the content of documents, as opposed to the frequently used approach of binary classification into fake and non-fake. Our classification system is based on analysis using graph neural networks, which allows for a more complex and contextual understanding of the data. The obtained results indicate a significant improvement in effectiveness compared to the baseline approaches, which suggests a potential for enhanced mitigation of misinformation dissemination.

1. Introduction

1.1. Context and rationale

In principle, nowadays, while looking at the current emerging problems of digital security one can distinguish several major challenges, namely: (i) cyber-attacks, (ii) security of artificial intelligence (AI) and (iii) spreading fake news (disinformation).

Cyber-attacks can be targeted at both critical national (or international) infrastructure such as, e.g., energy grids, telecommunication networks, hospitals, etc., as well as individual citizens and their data or the services they use. Similarly, AI systems can be targeted by so-called adversarial attacks on large scale (national systems) or smaller scale (personalized applications or services). Fake news and/or disinformation can also be recognized as a serious threat on large-scale and

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.
Email address: chorasm@pbs.edu.pl (M. Choraś).

https://doi.org/10.1016/j.neucom.2025.131327
Received 12 June 2025; Received in revised form 30 July 2025; Accepted 17 August 2025
Available online 19 August 2025

smaller-scale levels. As for the large-scale, well-crafted disinformation campaigns can not only influence elections but also divide societies, causing confusion, fear, panic and disagreements. As for the smaller scale, fake news can impact customer decisions or influence how people spend their free time etc. It is, however, important to acknowledge here that small-scale disinformation targeted at personalized channels or information bubbles can quickly emerge into large-scale societal problems and even a threat to national matters and security.

In this paper, we propose a novel, non-binary approach to digital security assessment in the context of disinformation detection. Our framework is based on a set of 13 diagnostic questions designed to capture the nuanced and multi-dimensional nature of information credibility. To classify the responses to these questions, we introduce a method grounded in graph neural networks (GNNs), which enables contextual and structurally informed reasoning. This approach allows for more flexible and accurate evaluation of disinformation across different levels of scale and personalization, bridging the gap between qualitative assessments and automated detection.

1.2. Contributions and structure of the paper

This paper is structured as follows. It begins with a literature review to provide context and identify existing approaches to addressing the issue of fake news. Following this, the proposed solution is presented, which highlights the methodology and its implementation-related aspects. The subsequent section focuses on experiments. We focus specifically on the SWAROG dataset, along with the analysis of the experiments conducted and the results obtained. Finally, the paper concludes with a summary of ideas for future work.

This paper makes the following key contributions:

- We propose a novel, non-binary framework for disinformation detection based on 13 diagnostic questions. Although these questions were previously introduced in other publications, our work extends their application by integrating them into a novel framework for disinformation detection.
- We introduce a graph neural network-based method for classifying the responses to these questions, which allows for more nuanced and accurate assessments of information credibility.
- We demonstrate the applicability of the proposed method using the SWAROG dataset, providing a detailed analysis of the experiments and results.

2. Related works

Text feature extraction is a key step in disinformation detection and more broadly in natural language processing tasks [1]. In the context of disinformation analysis, feature extraction involves extracting representative information from text that can be used to assess the truthfulness, intent, or style of the message contained within it. The next step is to classify the text, articles, social media posts, and identify relationships to better detect disinformation. All of these processes, in turn, pose a serious challenge in the context of ensuring true and broadly understood cybersecurity and public safety [2]. Researchers around the world are engaged in recognizing disinformation, clickbait, and rumors in native languages, as it is considered a significant problem [3]. The Polish language belongs to the largest family of Indo-European languages in terms of speakers. The table below presents an overview of the research on selected Indo-European languages to date [4]. For broader context, analyses of Arabic and multilingual studies are also included. A detailed discussion is provided in Table 1.

Among the feature extraction techniques, we can distinguish the very basic ones such as: N-gram [5] for modeling sequences of words or characters in text, bag-of-words [6] where the text is represented by a set of words, the TF method for calculating the frequency of a word in the text [7] or TF-IDF [8] as an extended method where the meaning and importance of a word in the text are also assessed. More advanced methods are also used, including transformers such as BERT [9], RoBERTa [10], and DistilBERT [11]. Hybrid feature extraction methods are also known, characterized by increased accuracy Kątek et al. [12].

Many of the works presented in the table focus on using different machine learning and deep learning models to improve the accuracy of detecting disinformation in news. Dinu, Fusu, and Gifu [20] presented an approach to detect fake news in Romanian using SVM and LR classifiers and text processing techniques such as TF-IDF and 300-dimensional CoRoLa vector embeddings. The best results were obtained using SVM and LR. In a similar vein, Valeanu et al. [21] presented a method for detecting vaccination messages on Twitter using SVM, MLP, and RF. Their model achieved AUC scores ranging from 0.744 to 0.858, indicating the high efficiency of the classification algorithms. Bucos and Ţucudean [19] used the Veridica dataset of Romanian news articles to analyze fake news. In their study, classifiers such as Extra Trees, RF, and SVM were used, achieving the best results using Back Translation technique. Moisi, Ţucudean, and Ionescu [23] presented an approach using BiLSTM and RoBERTa-large for detecting fake news in Romanian, achieving 96.5 % accuracy for the BERT-based model.

In the context of Arabic language, Al Ghamdi et al. [27] collected data from Twitter and articles, creating a corpus that includes both fake and real news. They analyzed different classification models such as Naive Bayes, Logistic Regression, SVM, and BERT, achieving 90 % accuracy using the BERT model. Sorour and Abdelkader [29] used a hybrid approach using CNN and LSTM for detecting fake news in Arabic, achieving 81 % accuracy.

Harris, Hadi, Ahmad et al. [28] developed a fake news detection model for Urdu using UrduFake@FIRE2020 datasets that contained both real and fake news from five domains. They used models such as ELECTRA, mBERT, and XLM-RoBERTa, achieving an accuracy of 91.4 % using an ensemble method.

In Azzeh, Qusef, and Alabboushi [30], the authors used six text representation techniques and five deep word embedding models such as AraBERT, AraELECTRA, ARBERT, MARBERT, and CAMeLBERT to detect fake news in Arabic. They achieved the best results using CAMeLBERT combined with a deep neural network (DNN), achieving an F1 of 71.3 % and an AUC of 79.1 %. Although many studies focus on individual languages, other works, such as Mohawesh, Maqsood, and Althebyan [32], consider multiple languages such as English, Hindi, Swahili, Vietnamese, and Indonesian. Their approach was based on capsule neural networks, which detected fake news with an improvement of about 3.97 % over the baseline models.

In the context of English, Keya et al. [34] used BERT embeddings combined with deep CNN and LSTM in the FakeStack model, which achieved an accuracy of over 97 % in detecting fake news. Han, Karunasekera, and Leckie [35] used graph neural networks to analyze the propagation patterns of fake news in a social network. Their approach, which took into account Twitter data and user characteristics, achieved an accuracy of 83 %. Among the latest works in this field, it is worth mentioning the research of Roy et al. [37], who created a new method for detecting fake news in Bengali, achieving an accuracy of 99 % using the Bidirectional Gated Recurrent Unit (GRU) model. Malla and Banka [38] presented an approach using Graph Attention Networks (GAT), which allowed them to take into account user preferences and social context, achieving an accuracy of 98 %. Other innovative approaches include the work of Frisli [39], who applied a semi-supervised self-training approach to the classification of disinformation, achieving an accuracy of over 98 %. In turn, Wanda and Diqi [40] in their study on the Indonesian language used a novel Generative Round Networks (GRN) architecture, achieving an accuracy of 94.33 %. Finally, in the context of the Albanian language, Canhasi et al. [41] used classical classification techniques such as KNN, XGBoost, and BERT and FastText embeddings to detect fake news. Their results showed the effectiveness of the classification methods in this language. On the other hand, for Indonesian language, Isa, Nico, and Permana [42] used the IndoBERT model, which achieved an accuracy of 94.66 % in detecting fake news,

Table 1
Summary of datasets and technologies for fake news detection in various languages.

Authors	Language	Data	Technology
Martínez-Gallego et al. [13]	Spanish	Spanish Fake News Corpus (971 news) + "Fake news in Spanish" (1600 news), totaling 2571 samples	BETO + LSTM
Blanco-Fernández et al. [14]	Spanish	A synthetic corpus of 57,231 political articles (46,000 - training, 11,231 - test) - data obtained via web scraping and generative models	Fine-tuned BERT / RoBERTa
Ibañez-Lissen et al. [15]	Spanish	News datasets (including social media data) - including Spanish fake news (e.g., related to the political situation in Spain)	GCN + BERT
Moreno-Vallejo et al. [16]	Spanish	Datasets obtained from social media and news articles	Compared MLP, CNN, and LSTM (LSTM achieved the best result)
Catelli et al. [17]	Italian	A new dataset related to Italian cultural heritage, containing reviews in Italian	BERT + ELECTRA + sentiment analysis
Buzea et al. [18]	Romanian	Online news dataset in Romanian	SVM, NB, LSTM, CNN, GRU, RoBERT-small, RoBERT-large
Bucos et al. [19]	Romanian	Dataset from the Romanian fact-checking website Factual.ro	Two data augmentation techniques: Back Translation and Easy Data Augmentation
Dinu et al. [20]	Romanian	Romanian news dataset from online sources	Compared models: LR, SVM, RF, SD Classifier, Dummy classifier
Valeanu et al. [21]	Romanian	Information related to vaccines in Romanian-language tweets	SVM, MLP, RF, Ensemble (SVM + MLP), RCNN, BERT
Daria-Mihaela et al. [22]	Romanian	Clickbait detection in Romanian-language news articles. The RoCliCo dataset (8,313 samples)	RF, SVM, BiLSTM, Fine-tuned Ro-BERT, Contrastive Ro-BERT, Ensemble
Moisi et al. [23]	Romanian	FakeRom - 1000+ articles collected through systematic scraping from the Veridica platform	Naive Bayes, Logistic Regression, SVM, BERT
Farooq et al. [24]	Urdu	From nine different domains, 4097 news, manually annotated	TF-IDF, BoW, Ensemble (Random Forest + Extra Trees), SVM, k-NN
Iqbal et al. [25]	Urdu	Tweets, fake and real labels, 12,047 posts	Feature extraction: frequency, inverse document frequency, SVM, Random Forest, Logistic Regression, Naive Bayes, Decision Tree, CNN, RNN
Munir et al. [26]	Urdu	Text + images, fake or not fake labels	-
Al Ghamdi et al. [27]	English, Arabic, Urdu	Twitter, web-based articles	Tokenization, Lemmatization, TF-IDF, Logistic Regression, Multinomial Naïve Bayes, Gradient Boosting, Decision Tree, Random Forest, KNN, BERT
Harris et al. [28]	Urdu	UrduFake@FIRE2020: 750 real news, 550 fake news from 5 domains	ELECTRA, mBERT, XLM-RoBERTa
Sorour et al. [29]	Arabic	1,475 Real, 3,152 Fake news	CNN + LSTM
Azzeh et al. [30]	Arabic	Combined dataset - websites and Twitter posts collected by the authors	AraBERT, AraELECTRA, ARBERT, MARBERT, CAMeLBERT, DNN, SVM, Logistic Regression, XGB, Naïve Bayes
E. Almandouh et al. [31]	Arabic	Publicly available Arabic datasets, 228,461 samples	FastText, Naïve Bayes, Logistic Regression, Linear SVC, Random Forest, SVM, Decision Trees, Gradient Boosting, XGB, CatBoost, AdaBoost
Mohawesh et al. [32]	Multiple languages	Multilanguage: English-English, English-Hindi, English-Indonesian, English-Swahili, English-Vietnamese [33]	Capsule neural network, mBERT, XLM, XLM-RoBERTa, MGL
Keya et al. [34]	English	WelFake, LIAR	BERT, FakeStack (BERT, deep CNN, LSTM)
Han et al. [35]	English	FakeNewsNet: labelled news from politifact.com and gossipcop.com	Graph-based neural networks
Lu et al. [36]	English	Tweets: Twitter15 and Twitter16	Graph-aware Co-Attention Network
Roy et al. [37]	Bengali	BanFakeNews dataset	Bidirectional Gated Recurrent Unit
Malla et al. [38]	English	FakeNewsNet, Gossip, Tweets	Graph Attention Networks (GAT)
Frisli, S [39]	Norwegian	426,262 tweets related to COVID-19, approximately 5.11 % as misinformation	Semi-supervised self-learning classifier using logistic regression with class weights
Wanda et al. [40]	Indonesian	Indonesian "Fake News"	Generative Round Networks (GRN)
Canhasi et al. [41]	Albanian	Dataset containing labelled true and fake news in Albanian	Logistic Regression, Naive Bayes, SVM, Decision Trees, Random Forest, KNN, XGBoost
Isa et al. [42]	Indonesian	COVID-19 news dataset in Indonesian	IndoBERT

especially those related to COVID-19. Heterogeneous graphs are gaining popularity in the context of disinformation analysis, which can extend the capabilities of large language models (LLMs) by integrating different types of entities and relationships. This integration enables more nuanced understanding and processing of complex data structures, which is essential for various applications in NLP and graph-based learning. Xie et al. [43] used a heterogeneous graph containing news, entity and topic nodes to model news content. The knowledge from the three knowledge graphs is then combined to extend the factual basis of news articles. Kang et al. [44] used a heterogeneous graph approach to exploit various relationships between news items, such as their temporal context, content, topic and source, to identify fake news. In this paper, they proposed building a heterogeneous graph, called News Detection Graph (NDG), containing different types of nodes and edges, which allows for the integration of multi-faceted data from multiple news items. Sun et al. [45] introduced a fake news detection model based on topic perception, where a given news article is divided into sentences and a heterogeneous graph is created from nodes representing sentences, topics and entities. Then, feature extraction and entity comparison are performed to assess semantic consistency. The use of graph neural networks is also popular [46]. Karnyoto et al. [47] in their work used heterogeneous graph neural network to detect disinformation about the COVID-19 pandemic. To create a graph neural network, they built nodes and edges in the graph as well as word-to-word and word-to-document nodes.

All of this work shows how a variety of methods and approaches can be effective in detecting fake news, including both classical techniques and modern deep learning models that are able to achieve impressive results across different languages and contexts.

In summary, the methods described in Table 1 exhibit both similarities and differences. Most of them are based on classification models such as SVM, RF, BERT, or their variants, but differ in text representation (TF-IDF, word embeddings, language models), analysis language, and context (e.g., social data, content propagation, graphs). The common goal of all approaches is to improve disinformation detection performance, yet the techniques employed are often limited to a single type of information (textual or structural). In contrast, the approach proposed in this paper combines a graph neural network with a classic TF-IDF representation and a classifier, allowing for the integration of both semantic and structural dependencies between messages. This allows for the capture of deeper relationships between content, resulting in a significant improvement in performance compared to existing methods.

3. Proposed approach

In this section, an overview of the proposed approach is provided. First, the explanation of the data normalization process is commenced. Next, the feature extraction and classification matters are explained.

3.1. Data normalization and cleaning

The data cleaning procedure used in this study was a two-stage process. The first stage was a stage of preliminary cleaning of the text so that it contained only words and basic punctuation marks such as periods, commas, question marks, and exclamation marks. In order to complete this stage, special characters, brackets, numbers, and HTML tags were removed from the data. In this step, empty and duplicate records with contradictory annotations were also removed.

The removal of numerical data during the text cleaning process was a deliberate and justified decision aimed at improving the model's generalization ability and reducing noise in the dataset. In many cases, numbers, especially when taken out of context, do not contribute significantly to understanding the overall semantic content of the message. Moreover, numerical values are often highly context-dependent and can appear in both real and fake news texts in similar forms, making them less reliable as standalone indicators of veracity. Removing such data helped focus the model on the linguistic and rhetorical patterns that are more generalizable across different texts and more indicative of fake news, such as sensational language, emotional tone, or misleading phrasing.

The second stage of the text cleaning process was to remove insignificant words (stop words). The decision to remove them was made by the authors of the article because of the need to better focus on words that carry greater meaning and have a direct impact on the meaning of the text. In addition, in this stage, the labels true and false were converted to their corresponding binary values.

After completion of this stage, a dataset of 3986 records was obtained.

3.2. Feature extraction

Two feature extraction methods were used in this study. The first one was based on the use of the DistilBERT model, pre-trained for the Polish language [48]. This model is characterized by its lightness and increased efficiency compared to the original BERT model [11].

The texts of the articles cleaned in the previous stage were first submitted to the tokenization process. This process consisted of adding unique tokens [CLS] and [SEP] at the beginning and end of the text. The next step was to replace all words with unique identifiers from the embedding table corresponding to the selected model. In the last step, the text prepared in this way was transferred to the DistilBERT model to generate an embedding vector. Finally, after this stage, a feature vector of 768 elements was obtained.

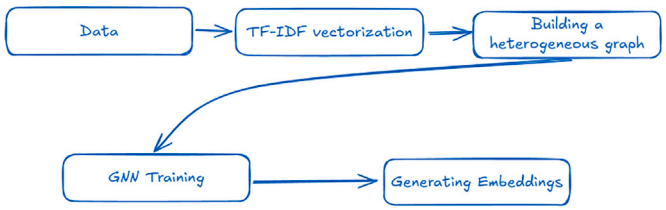

Fig. 1. Flow chart of the GNN-based method.

The hybrid method based on the heterogeneous graph network is the second method developed in this paper. Basically, this method combines the TF-IDF text vectorization technique with heterogeneous graph modeling. However, this technique consists of several stages, which are presented in detail in Fig. 1.

In the first step, text data (including authors, titles, and content of articles) are transformed into a numerical representation using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This process includes several important steps, which are presented in the form of pseudocode below (Algorithm 1).

The formulas to calculate the values of the values of TF (formula 1) and IDF (formula 2) values are defined below.

$$TF(t,d) = \frac{\text{number of occurrences of word } t \text{ in text } d}{\text{total number of words in document } d} \tag{1}$$

$$IDF(t) = \log\left(\frac{\text{number of sentences in article } t}{\text{number of sentences containing word } t + 1}\right) \tag{2}$$

In the next stage, after the text data has been transformed into vector form, it is mapped into a heterogeneous graph, in which different types of data (content, authors, and titles) are transformed into their corresponding nodes. The relations between them have been defined in the form of graph edges, e.g., the title is related to the content, or the author is related to the content he wrote. In addition, the authors introduced relations based on semantic similarity between the title and content nodes to the graph structure. For this purpose, the cosine similarity metric was used, calculated on the basis of the feature vectors obtained in the previous stage, and defined by the formula below 3.

$$w(u,v) = \cos(h_u, h_v) = \frac{h_u \cdot h_v}{\|h_u\| \|h_v\|} \tag{3}$$

where:

- h_u, h_v - node feature vectors (title and content embeddings),
- $\cos(h_u, h_v)$ - cosine similarity measure defining the strength of the relationship between nodes.

In this work, TF-IDF was used as the method for vectorizing textual content. Each title and article content were transformed into a fixed-size

Algorithm 1 Text processing using TF-IDF.

Input: Document collection D
Output: Vector representation of documents
for each document $d \in D$ **do**
    Remove stop-words and punctuation marks from d
    for each word t in d **do**
        Calculate the frequency of the term $TF(t,d)$
        Calculate the rarity of $IDF(t)$ in the collection D
        Calculate the value $TF - IDF(t,d) = TF(t,d) \times IDF(t)$
    end for
    Create a vector representation V_d of document d
end for
return A collection of vectors representing documents

TF-IDF vector representation. These vectors were then assigned to the corresponding nodes in the graph.

In the final stage of the data processing pipeline, Graph Neural Networks were employed to fully exploit the graph structure obtained in the previous phase. The application of GNNs enables not only the modeling of relationships between entities but also the extraction of latent patterns and semantic dependencies embedded in the textual data. The resulting rich node representations (embeddings) serve as the foundation for the subsequent information propagation process within the graph.

In contrast to traditional homogeneous graphs, the present case involves a heterogeneous graph where nodes can represent distinct types of entities, such as *author*, *title*, and *content* and where edges denote semantically different relations. For instance, the `wrote` edge links an author to content, while the `title-of` edge connects a title to its associated content. This structural diversity necessitates specialized handling at the model level, whereby separate layers are applied for each node and edge type pair. This architectural choice allows independent learning of weights for different relation types, thereby enhancing the model's capacity to distinguish between contextual and semantic meanings of connections.

The graphical diagram in Fig. 2 illustrates a simplified structure of such a graph: it includes two authors (`Author 1`, `Author 2`), two titles (`Title 1`, `Title 2`), and their corresponding contents (`Content 1`, `Content 2`). These nodes are interconnected by logical relations such as `wrote` and `title-of`, reflecting common linkages present in natural documents.

Moreover, the graph includes semantic relations like `CosineSim` between titles and contents, as well as general similarity relations (`Similar`) between title and content pairs. These edges are derived from vector-based textual similarity measures (cosine similarity), enabling the model to capture thematic proximity regardless of structural links. As a result, the system can recognize content with related topics, even if written by different authors or expressed in varied linguistic styles.

In the adopted GNN architecture, each node type is assigned a type-specific weight matrix, responsible for transforming node features in a contextually appropriate way. This allows for differentiated information propagation, depending on both the node's type and the types of its neighbors. Information propagation itself is iterative: in each layer (iteration), a node's representation is updated by aggregating information from its neighbors, weighted appropriately.

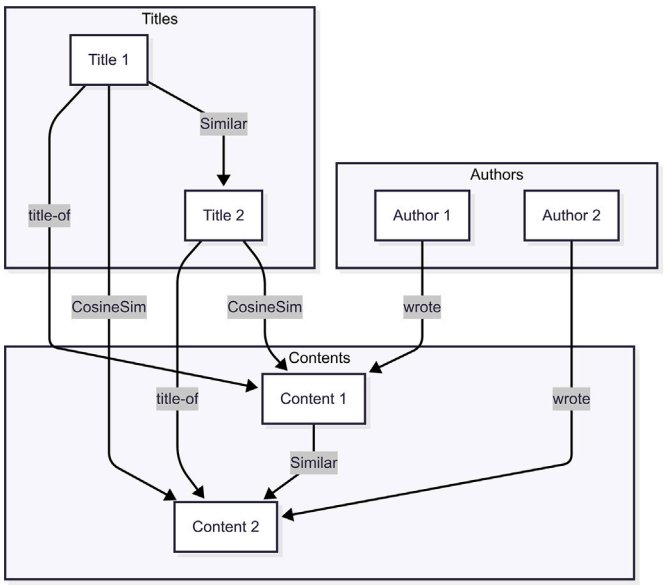

Fig. 2. Structure of a heterogeneous graph with logical and semantic relations between authors, titles, and contents.

After several rounds of message passing, final node embeddings are produced. These embeddings encode both the structural information (who is connected to whom and how) and the semantic similarity of nodes. They are then used to classify answers to questions proposed by the authors.

Such an architecture allows not only for precise modeling of dependencies in textual data but also for dynamic adaptation to its contextual complexity. In particular, the integration of semantic similarity relations with logical graph structures constitutes a significant extension to traditional GNN approaches, enhancing the model's flexibility and performance in real-world applications.

The training of the model was carried out using the Adam optimizer with a learning rate of 0.01 over 30 epochs. During each epoch, the node representations were updated through two layers of heterogeneous graph convolutional networks. These modules utilized the GraphSAGE aggregation mechanism, where the node representations were updated by aggregating information from neighboring nodes according to the edge types. The training process involved classifying the node embeddings, using the cross-entropy loss function, and optimizing the model parameters through backpropagation. The model was trained in a full-batch fashion, meaning all available training data were used at each iteration without node masking.

An ablation study was conducted to assess the impact of different model components on performance. Removing the semantic similarity edges led to a significant decrease in classification accuracy, indicating the crucial role these edges play in capturing thematic proximity between nodes.

Additionally, when the graph was simplified to a homogeneous structure, where all nodes and edges were treated uniformly, the model's performance dropped by approximately 15 %, highlighting the importance of handling heterogeneous node and edge types. Finally, a comparison between the GraphSAGE aggregation mechanism and mean pooling demonstrated that the GraphSAGE-based model outperformed mean pooling by 10 %, underscoring the advantage of specialized message passing in preserving nuanced relationships. These results reinforce the value of the model's complex design and the integration of semantic relations in improving performance.

4. Experiments and results

4.1. Dataset used for the experiments

The SWAROG dataset is available online on GitHub[1] and has been described in our previous work [49].

The data is made available without an imposed division into cross-validation folds. Since the phenomenon of fake news is a multi-dimensional concept, the proposed approach is not limited to a one-dimensional binary response. Unlike other fake news datasets (which commonly use binary classification), we decompose the problem into several dimensions, namely: Sources and Credibility, Context and Precision, Author's Intentions, and Nature of the Content. For each of these factors, 13 well-defined questions commonly used by fact-checkers were proposed. The structure and type of the questions have been presented in Fig. 3.

To ensure a comprehensive evaluation of misinformation, the questions were grouped into three major categories, each representing a specific aspect of content credibility:

1. **Verification Factors (Sources and Credibility):** These questions assess whether the content presented in the article is supported by external, reliable sources. They require the annotator to verify the information by consulting additional references. This category includes the following questions:

[1] https://github.com/w4k2/swarog-dataset.

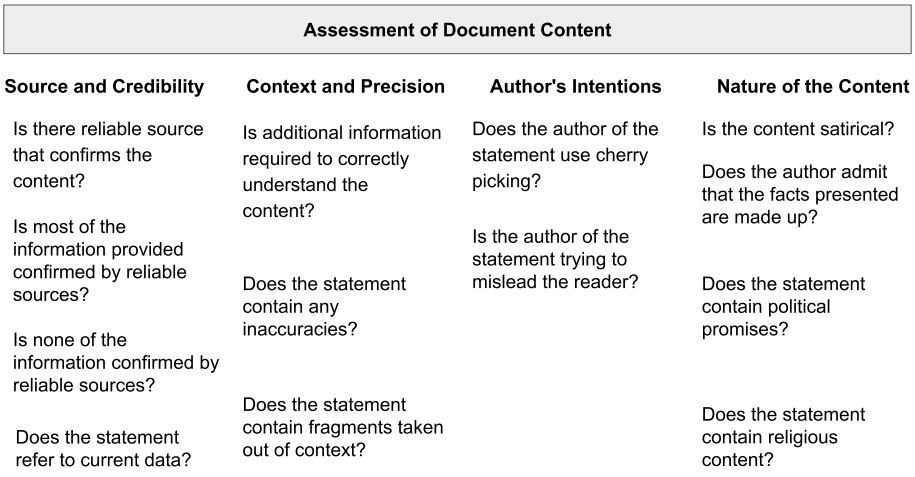

Fig. 3. Factors and questions used for assessing the document content.

- Is there at least one reliable source that confirms all the information contained in the content?
- Is most of the information provided confirmed by reliable sources?
- Is none of the information confirmed by reliable sources?
- Does the statement refer to current data?

These questions help evaluate the **objectivity** and **factual accuracy** of the content.

2. **Manipulative Factors (Author's Intentions and Context):** This group focuses on identifying indicators of manipulation or misleading intent. The goal is to detect deliberate attempts to distort facts or present information in a biased or incomplete way. The following questions are used:
   - Is additional information required to properly understand the content?
   - Does the content contain inaccuracies?
   - Does the statement contain fragments taken out of context?
   - Does the author of the statement use cherry-picking?
   - Is the author of the statement trying to mislead the reader?

   These questions aim to uncover whether the **intent** behind the content is to influence the reader deceptively or push a particular agenda.
3. **Metaphysical Factors (Nature of the Content):** These questions examine the **tone** and **narrative style** of the content, which may influence the emotional reaction of the reader. Authors often tailor content to resonate with the beliefs or values of their target audience, thereby increasing its reach and impact. The questions include:
   - Is the content satirical?
   - Does the author admit that the facts presented are made up?
   - Does the statement contain political promises?
   - Does the statement contain religious content?

   These questions help assess whether the content is intended to be taken literally or figuratively, and whether it relies on **ideological**, **religious**, or **emotional appeal**.

All questions were precisely formulated and designed to allow binary responses (*"yes"* or *"no"*) by the annotators. This consistent structure enhances the reliability and comparability of the annotations and supports more nuanced, multi-label modeling approaches beyond simple true/false classification.

The SWAROG dataset is diverse and balanced in terms of the subjects and content. This was achieved through adequate sampling of news that appeared in the public domain, which was later passed to annotators [49]. The t-SNE plot shows that the clusters are distinct, reflecting the diversity of the content. The dataset includes news on topics such as the war in Ukraine, politics, sports, and celebrity gossip Fig. 4.

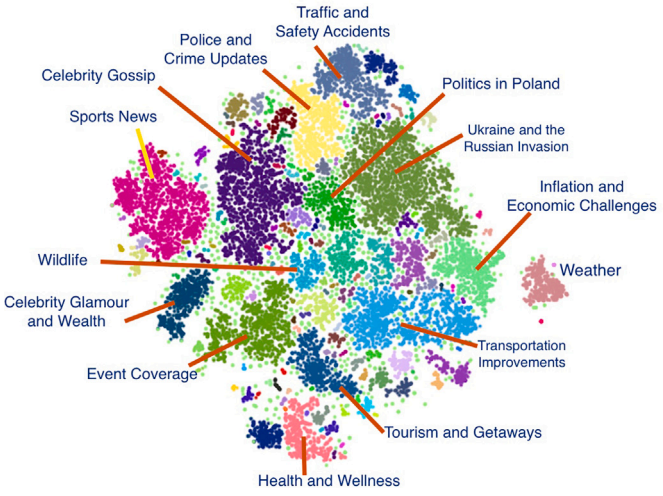

Fig. 4. t-SNE visualization of the SWAROG dataset: We used S-BERT to encode the news titles.

4.2. Results and discussion

The experiments presented used the ten-fold cross-validation technique. The graphs (graphs from 5 to 8) show the results of accuracy, balanced accuracy, and f1 (0) and f1 (1), where 0 is labeled false and 1 is true for each of the questions using the proposed feature extraction methods. The experiments presented used the same dataset for both extraction methods, which was balanced by randomly rejecting samples. Both feature extraction methods were tested using the decision tree from the scikit-learn library with default parameters.

In the factor group, the highest results were obtained for the question, 'Is most of the information provided confirmed by reliable sources?' where the highest efficiency was 0.94. The results of the remaining metrics for this question achieve similar results (Fig. 5).

For the question, 'Is there at least one reliable source that confirms all the information contained in the content?' the highest efficiency of the graph-based method was 0.89 and the DistilBERT method was 0.61. The results of the other metrics for this question achieved similar results (Fig. 6).

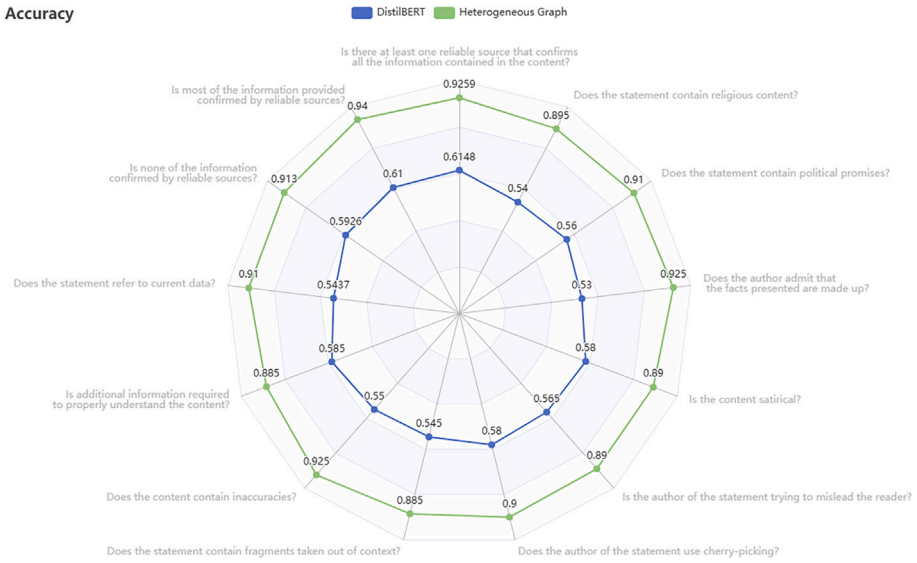

Fig. 5. Experiment results for individual questions- accuracy metric.

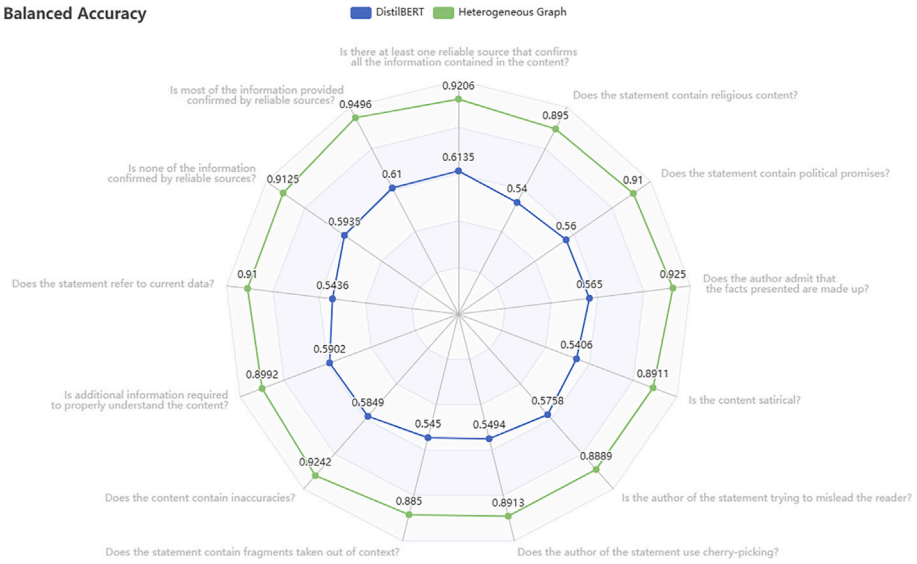

Fig. 6. Experiment results for individual questions- balanced accuracy metric.

In the case of the question 'Is none of the information confirmed by reliable sources?' the highest efficiency was 0.93 for the graph-based method, while for the DistilBERT method the efficiency was 0.59. The results of the other metrics for this question achieved similar results (Fig. 7).

In the question 'Does the statement refer to current data?', the efficiency of the graph-based method was 0.91, and the BERT-based method was 0.54. The results of the other metrics for this question achieved similar results (Fig. 8).

The next group of factors is manipulative factors. The results for the question are as follows. 'Is additional information required to properly understand the content?', the effectiveness of the graph method reached a value of 0.89, while that based on transformers reached 0.58, which is presented in the graph.

For the next question in this group 'Does the content contain inaccuracies?', the effectiveness of the proposed method reached a value of 0.93.

In turn, in the question regarding the author's use of the cherry-picking method, the effectiveness of the BERT method reached a value of 0.55, while the method based on heterogeneous graphs reached an effectiveness of 0.9.

The last question in this group is a question about misleading the reader, where the graph method reached an effectiveness of 0.89, while the effectiveness of the method based on transformers was 0.58.

The last group of factors is metaphysical factors. The first question in this group regarding the satirical nature of the article was assessed with an efficiency of 0.89 for feature extraction based on the graph method and 0.56 for the DistilBERT method.

In the next question regarding the author's admission that the presented content is made up, the graph method managed to achieve an efficiency of 0.93, and the BERT-based method at 0.53.

The question regarding political promises achieved a result of 0.91, with a result of 0.56 for the DistilBERT method.

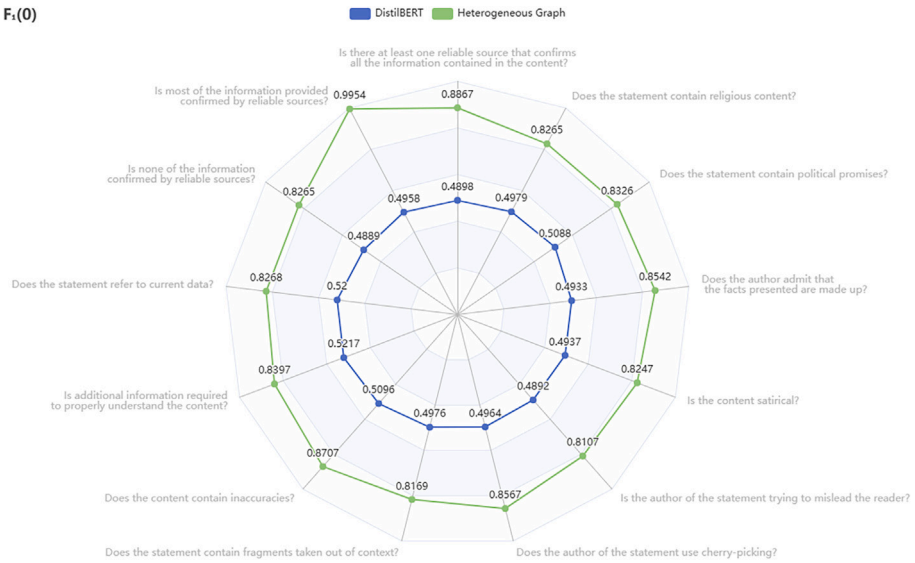

Fig. 7. Experiment results for individual questions- F1 (0) metric.

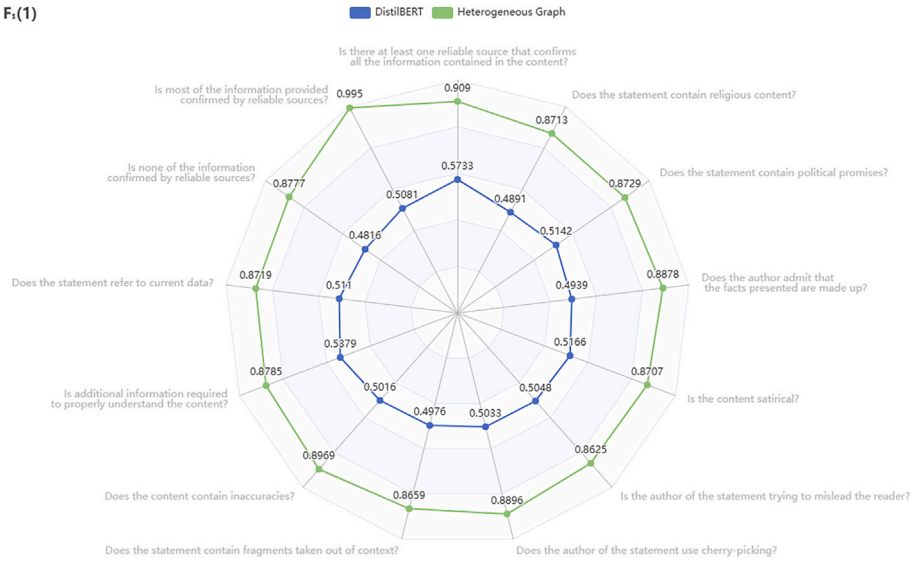

Fig. 8. Experiment results for individual questions- F1 (1) metric.

For the question regarding the content of fragments taken out of context, the proposed method using heterogeneous graphs achieved an efficiency of 0.89, while the method based on transformers achieved an efficiency of 0.55.

In the last question about religious content, the BERT-based method achieved an efficiency of 0.54, and the heterogeneous graph-based method 0.9.

The analysis of the results presented in the article clearly indicates a clear advantage of methods based on heterogeneous graphs over the approach based on transformer models in the tasks of assessing the credibility of information and identifying manipulations in the content. In particular, for the question of confirming the majority of the information provided in credible sources, the proposed solution achieved the highest effectiveness of 0.94, which confirms that the graph approach effectively identifies connections between sources and content.

A similar trend was observed in the question regarding the full compliance of the content with a credible source, where the graph method achieved 0.89, and the transformer approach only 0.61.

In the task of identifying the lack of confirmation of information in credible sources, differences in the effectiveness of the proposed methods are again visible. These differences result from the fact that graph models can better reflect the structural connections between information units, which is a key issue when analyzing the consistency and credibility of content.

The authors noticed a similar situation in the case of analysis of manipulation factors. In the question regarding the need to obtain additional information to fully understand its content, the graph model's effectiveness was 0.89, and the transformer model's 0.58. Importantly, in the assessment of the presence of cherry picking, the graph model achieved a result of 0.9, significantly outperforming the BERT method. Moreover, in the question regarding deliberately misleading the reader,

the graph method achieved an effectiveness of 0.89, while the transformer method only 0.58. These results indicate that graph models can better capture structural dependencies in text, which are important for detecting manipulation.

The last group of questions analyzed in the article is the group of metaphysical factors. In this group, similarly to the other groups, a clear advantage of graph methods was observed.

To sum up, the analysis of the obtained results confirms that the graph approach offers better effectiveness in tasks related to assessing the credibility of information and identifying manipulation in the content, both in the context of factual, manipulative, and metaphysical factors. In the future, the authors plan to extend the research proposed in this paper with a hybrid approach that allows to extract the best features of the graph-based method with the best features of the DistilBERT models to further improve the obtained results and be able to detect more complex text manipulations.

5. Threats to validity

In this article, an innovative and original methodology to assess textual information based on a novel methodology of 13 questions was presented, rather than the typical binary true-false approach.

The authors of this work believe that disinformation detection should not be based on binary decisions (true-false), however, the authors fully understand that these 13 questions could be extended or modified depending on the culture or specific domains as well as languages.

For example, the question about religious content in the text is an indicator of the culture of the authors, but the authors are fully aware that there are cultures where everything is connected to dominant religion, or that there are secular societies where religion is non-factor in disinformation detection.

In practical terms, the authors of the study hired an independent company and annotators to assess selected real texts with information. Of course, the authors tracked their response times, etc. to eliminate noise and low-quality annotations. Moreover, the authors collected information on annotators in order to avoid bias and track their backgrounds. However, it is obvious that some annotations could be influenced by fatigue or other behavioral aspects.

Another specificity of the presented work is the analysis of information in Polish language. Still, the authors believe that, in principle, both the approach and technical methods can be generalized to other languages as well. The authors are also aware that the dataset and approach target the domain of general news. The results might be worse if the analyzed sources (trained on the presented dataset, approach, and model) are from another specific and narrow domain.

6. Conclusions

Fake news poses a significant threat to state security and the functioning of public spaces, particularly as the volume of information shared by news services and social media grows exponentially. This makes manual verification of all news items or press releases practically unfeasible.

This article introduces an innovative approach to combating fake news by employing a multi-factor assessment of document content rather than the conventional binary classification into fake and non-fake categories.

The proposed method leverages heterogeneous graphs for data representation and analysis, offering a novel perspective on feature extraction.

Experimental results demonstrate that the graph-based method consistently outperformed the widely used ‘vanilla’ DistilBERT approach across all question groups and evaluation metrics, highlighting its effectiveness in this domain. The experiments were rigorously validated using a ten-fold cross-validation approach, ensuring robust and reliable results.

This paper builds upon our previous work [50] and introduces several key advancements. While both studies utilize the same 13-question assessment, the current work replaces the earlier transformer-based approach with a heterogeneous graph neural network architecture. This methodological shift enables the model to better capture the structural and semantic relationships inherent in the dataset, resulting in improved performance across all evaluation metrics. Furthermore, this paper places greater emphasis on interpretability and robustness through detailed ablation studies and cross-validation, which were not the focus of the previous publication. These contributions mark a significant step forward in the development of explainable and high-performing models for content credibility assessment.

In conclusion, the proposed multi-factor approach represents a significant advancement in content credibility assessment, moving beyond the traditional binary classification of “fake” and “non-fake” categories. While the model generates richer, multi-dimensional outputs, it can also be transformed into a binary format by applying appropriate thresholds, weights, or aggregation functions. This flexibility allows users or researchers to tailor the results to specific needs or applications, depending on the level of uncertainty they are willing to tolerate. As such, our approach can not only expand upon existing methods but also be directly compared with them when necessary.

The future work will focus on the problem of concept drift. Over time, the SWAROG dataset undergoes aging, which can impact its relevance and usability. A critical aspect of this research will be detecting the occurrence of concept drift and evaluating the level of degradation in classifier performance caused by this phenomenon. We are also fully aware of the dual nature of transparency in multi-factor assessment models. While it enhances interpretability and trust, it may also expose the system to adversarial manipulation. As such, future research will explore protective strategies that safeguard explainable components of the model without compromising their utility, aiming to strike a balance between transparency and robustness in real-world applications. In our related work [51] we explored this issue in more detail, highlighting specific threats and mitigation strategies linked to the interpretability of misinformation detection systems.

CRediT authorship contribution statement

Gracjan Kątek: Writing – original draft, Validation, Software, Investigation, Formal analysis. **Rafał Kozik:** Writing – original draft, Visualization, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Aleksandra Pawlicka:** Writing – original draft, Validation, Methodology. **Marek Pawlicki:** Writing – original draft, Validation, Investigation, Formal analysis. **Michał Choraś:** Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Gracjan Kątek reports that financial support was provided by the National Center for Research and Development within INFOSTRATEG program. Rafał Kozik reports that financial support was provided by the National Center for Research and Development within INFOSTRATEG program and by Horizon Europe. Aleksandra Pawlicka reports that financial support was provided by Horizon Europe (Starlight project). Marek Pawlicki reports that financial support was provided by the National Center for Research and Development within INFOSTRATEG program and by Horizon Europe and by Horizon Europe (Starlight project). Michał Choraś reports that financial support was provided by the National Center for Research and Development within INFOSTRATEG program and by Horizon Europe by Horizon Europe (Starlight project).

Acknowledgement

This publication is partially funded by the National Center for Research and Development within INFOSTRATEG program, number of application for funding: INFOSTRATEG-I/0019/2021–00.

The work described in this paper is also partially performed in the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101021797.

We would also like to thank the KSSK group (https://kssk.pwr.edu.pl) from Wrocław University of Technology for developing and providing the dataset.

Data availability

Data will be made available on request.

References

[1] A. Tabassum, R.R. Patil, A survey on text pre-processing & feature extraction techniques in natural language processing, Int. Res. J. Eng. Technol. 7 (6) (2020) 4864–4867.
[2] K.M. Caramancion, An exploration of disinformation as a cybersecurity threat, in: 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020, pp. 440–444, https://doi.org/10.1109/ICICT50521.2020.00076.
[3] X. Zhou, R. Zafarani, A survey of fake news: fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (5) (2021) 1–40.
[4] B. Comrie (Ed), the World's Major Languages, 3rd edition ed., Routledge, London, England, 2020.
[5] W.B. Cavnar, J.M. Trenkle, N-gram-based text categorization, https://dsacl3-2019.github.io/materials/CavnarTrenkle.pdf, Accessed: 2023-Dec-11.
[6] H. Jiang, Y. Xiao, W. Wang, Explaining a bag of words with hierarchical conceptual labels, World Wide Web 23 (3) (2020) 1693–1713.
[7] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval (1972), in: Ideas that Created the Future, The MIT Press, 2021, pp. 339–348.
[8] L. Shi, R.-L. Xu, Research on deep learning model based on word2vec and improved tf-Idf algorithm, Comput. Digit. Eng. 49 (5) (2021) 966–970.
[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805, (2018).
[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: a robustly optimized BERT pretraining approach, arXiv:1907.11692, (2019).
[11] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv:1910.01108, (2019).
[12] G. Kątek, M. Gackowska, J. Komorniczak, P. Ksieniewicz, R. Kozik, M. Pawlicki, M. Choraś, Involving Society to Protect Society from Fake News and Disinformation: Crowdsourced Datasets and Text Reliability Assessment, Springer Nature Singapore, Singapore, 2024, pp. 384–395.
[13] K. Martínez-Gallego, A.M. Álvarez-Ortiz, J.D. Arias-Londoño, Fake news detection in Spanish using deep learning techniques, arXiv:2110.06461, (2021).
[14] Y. Blanco-Fernández, J. Otero-Vizoso, A. Gil-Solla, J. García-Duque, Enhancing misinformation detection in Spanish language with deep learning: BERT and ROBERTA Transformer models, Appl. Sci. (Basel) 14 (21) (2024) 9729.
[15] L. Ibañez-Lissen, L. González-Manzano, J.M. de Fuentes, M. Goyanes, On the feasibility of predicting volumes of fake news—the spanish case, IEEE Trans. Comput. Soc. Syst. 11 (4) (2024) 5230–5240.
[16] P.X. Moreno-Vallejo, G.K. Bastidas-Guacho, P.R. Moreno-Costales, J.J. Chariguaman-Cuji, Fake news classification web service for spanish news by using artificial neural networks, Int. J. Adv. Comput. Sci. Appl. 14 (3) (2023) https://doi.org/10.14569/IJACSA.2023.0140334.
[17] R. Catelli, L. Bevilacqua, N. Mariniello, V.S. Di Carlo, M. Magaldi, H. Fujita, G. De Pietro, M. Esposito, A new Italian cultural heritage data set: detecting fake reviews with BERT and ELECTRA leveraging the sentiment, IEEE Access (2023) 1.
[18] M.C. Buzea, S. Trausan-Matu, T. Rebedea, Automatic fake news detection for Romanian online news, Information (Basel) 13 (3) (2022) 151.
[19] M. Bucos, G. Ţucudean, Text data augmentation techniques for fake news detection in the Romanian language, Appl. Sci. (Basel) 13 (13) (2023) 7389.
[20] L. Dinu, E.C. Fusu, D. Gifu, Veracity analysis of Romanian fake news, Procedia Comput. Sci. 225 (2023) 3303–3312.
[21] A. Valeanu, D.P. Mihai, C. Andrei, C. Puscasu, A.M. Ionica, M.I. Hinoveanu, V.P. Predoi, E. Bulancea, C. Chirita, S. Negres, C.D. Marineci, Identification, analysis and prediction of valid and false information related to vaccines from Romanian tweets, Front. Public Health 12 (2024) 1330801.
[22] R.A. Dar, R. Hashmy, A survey on Covid-19 related fake news detection using machine learning models, in: Momlet + ds, 2023, https://api.semanticscholar.org/CorpusID:259254278.
[23] E.V. Moisi, B.C. Mihalca, S.M. Coman, A.M. Pater, D.E. Popescu, Romanian fake news detection using machine learning and transformer-based approaches, Appl. Sci. (Basel) 14 (24) (2024) 11825.
[24] M.S. Farooq, A. Naseem, F. Rustam, I. Ashraf, Fake news detection in Urdu language using machine learning, PeerJ Comput. Sci. 9 (2023) e1353.
[25] Z. Iqbal, F.M. Khan, I.U. Khan, I.U. Khan, Fake news identification in urdu tweets using machine learning models, Asian Bull. Big Data Manag. 4 (1) (2024).
[26] S. Munir, M. Asif Naeem, Bil-Fand: leveraging ensemble technique for efficient bilingual fake news detection, Int. J. Mach. Learn. Cybern. 15 (9) (2024) 3927–3949.
[27] M.A. Al Ghamdi, M.S. Bhatti, A. Saeed, Z. Gillani, S.H. Almotiri, A fusion of BERT, machine learning and manual approach for fake news detection, Multimed. Tools Appl. 83 (10) (2023) 30095–30112.
[28] S. Harris, H.J. Hadi, N. Ahmad, M.A. Alshara, Multi-domain urdu fake news detection using pre-trained ensemble model, Sci. Rep. 15 (1) (2025) 8705.
[29] S.E. Sorour, H.E. Abdelkader, Afnd: arabic fake news detection with an ensemble deep CNN-LSTM model, J. Theor. Appl. Inf. Technol. 100 (2022).
[30] M. Azzeh, A. Qusef, O. Alabboushi, Arabic fake news detection in social media context using word embeddings and pre-trained transformers, Arab. J. Sci. Eng. 50 (2) (2025) 923–936.
[31] M.E. Almandouh, M.F. Alrahmawy, M. Eisa, M. Elhoseny, A.S. Tolba, Ensemble based high performance deep learning models for fake news detection, Sci. Rep. 14 (1) (2024) 26591.
[32] R. Mohawesh, S. Maqsood, Q. Althebyan, Multilingual deep learning framework for fake news detection using capsule neural network, J. Intell. Inf. Syst. 60 (3) (2023) 1–17.
[33] A. De, D. Bandyopadhyay, B. Gain, A. Ekbal, A transformer-based approach to multilingual fake news detection in low-resource languages, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22 (1) (2023) 1–20.
[34] A.J. Keya, H.H. Shajeeb, M.S. Rahman, M.F. Mridha, Fakestack: hierarchical tri-BERT-CNN-LSTM stacked model for effective fake news detection, PLoS One 18 (12) (2023) e0294701.
[35] Y. Han, S. Karunasekera, C. Leckie, Graph neural networks with continual learning for fake news detection from social media, arXiv:2007.03316, (2020).
[36] Y.-J. Lu, C.-T. Li, GCAN: graph-aware Co-attention networks for explainable fake news detection on social media, arXiv:2004.11648, (2020).
[37] U. Roy, M.S. Tahosin, M.M. Hasan, T. Islam, F. Imtiaz, M.R. Sadik, Y. Maleh, R.B. Sulaiman, M.S. Hassan Talukder, Enhancing Bangla fake news detection using bidirectional gated recurrent units and deep learning techniques, in: Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security, vol. 2020, ACM, New York, NY, USA, 2024, pp. 1–10.
[38] A.M. Malla, A.A. Banka, Sustainable signals: a heterogeneous graph neural framework for fake news detection, Int. J. Syst. Assur. Eng. Manag. (2024).
[39] S. Frisli, Semi-supervised self-training for Covid-19 misinformation detection: analyzing Twitter data and alternative news media on Norwegian Twitter, J. Comput. Soc. Sci. 8 (2) (2025).
[40] P. Wanda, M. Diqi, Deepnews: enhancing fake news detection using generative Round Network (GRN), Int. J. Inf. Technol. 16 (7) (2024) 4289–4298.
[41] E. Canhasi, R. Shijaku, E. Berisha, Albanian fake news detection, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21 (5) (2022) 1–24.
[42] S.M. Isa, G. Nico, M. Permana, INDOBERT for Indonesian fake news detection, ICIC Express Lett. 16 (3) (2022).
[43] B. Xie, X. Ma, X. Shan, A. Beheshti, J. Yang, H. Fan, J. Wu, Multiknowledge and LLM-inspired heterogeneous graph neural network for fake news detection, IEEE Trans. Comput. Soc. Syst. (2024) 1–13, https://doi.org/10.1109/TCSS.2024.3488191.
[44] Z. Kang, Y. Cao, Y. Shang, T. Liang, H. Tang, L. Tong, Fake News Detection with Heterogenous Deep Graph Convolutional Network, Springer International Publishing, Cham, 2021, pp. 408–420.
[45] L. Sun, H. Wang, Topic-aware fake news detection based on heterogeneous graph, IEEE Access 11 (2023) 103743–103752, https://doi.org/10.1109/ACCESS.2023.3318483.
[46] L. Xu, J. Peng, X. Jiang, E. Chen, B. Luo, Graph neural network based on graph kernel: a survey, Pattern Recognit. 161 (111307) (2025) 111307, https://doi.org/10.1016/j.patcog.2024.111307.
[47] A.S. Karnyoto, C. Sun, B. Liu, X. Wang, Augmentation and heterogeneous graph neural network for aaai2021-Covid-19 fake news detection, Int. J. Mach. Learn. Cybern. 13 (7) (2022) 2033–2043, https://doi.org/10.1007/s13042-021-01503-5.
[48] SDADAS/polish-distilroberta · hugging face, https://huggingface.co/sdadas/polish-distilroberta, Accessed: 2023-Dec-11.
[49] R. Kozik, J. Komorniczak, P. Ksieniewicz, A. Pawlicka, M. Pawlicki, M. Choraś, Swarog Project approach to fake news detection problem, in: Computational Intelligence in Security for Information Systems Conference, Springer, 2023, pp. 79–88.
[50] R. Kozik, G. Kątek, M. Gackowska, S. Kula, J. Komorniczak, P. Ksieniewicz, A. Pawlicka, M. Pawlicki, M. Choraś, Towards explainable fake news detection and automated content credibility assessment: Polish internet and digital media use-case, Neurocomputing 608 (128450) (2024) 128450.
[51] R. Kozik, M. Ficco, A. Pawlicka, M. Pawlicki, F. Palmieri, M. Choraś, When explainability turns into a threat - using XAI to fool a fake news detection method, Comput. Secur. 137 (103599) (2023) 103599.