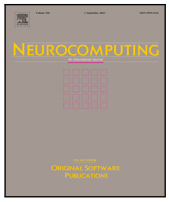
Neurocomputing 608 (2024) 128450



Contents lists available at ScienceDirect

Neurocomputing

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





Towards explainable fake news detection and automated content credibility assessment: Polish internet and digital media use-case

Rafał Kozik [a,f], Gracjan Kątek [a], Marta Gackowska [a], Sebastian Kula [a,c,e], Joanna Komorniczak [b], Paweł Ksieniewicz [b], Aleksandra Pawlicka [d,f], Marek Pawlicki [a,f], Michał Choraś [a,f,*]

[a] Bydgoszcz University of Science and Technology, Poland
[b] Wrocław University of Science and Technology, Wrocław, Poland
[c] Kazimierz Wielki University, Bydgoszcz, Poland
[d] University of Warsaw, Warsaw, Poland
[e] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
[f] ITTI Sp. z o.o., Poznań, Poland

ARTICLE INFO

Keywords:
Fake news
Machine learning
AI
Explainability

ABSTRACT

Detecting fake news and disinformation is a challenging and complex task. With the proliferation of the transformer architectures, researchers and practitioners have started actively using them to combat this phenomenon. In that regard, different approaches have been proposed by the community. Unfortunately, a significant number of solutions have started to treat misinformation in a similar way as other natural language processing challenges (e.g., sentiment classification). However, the reality is more complex, and the fake news detection problem often require additional context that may be essential to provide a complete assessment. Another pitfall in combating fake news is treating it as a binary classification challenge (e.g., fake vs. legitimate news). Combining such a simplified response with a deep black-box neural system effectively impacts the interpretability and eventual trustworthiness of the detection system. In this paper, the phenomenon of fake news is approached from a multi-factor perspective. Therefore, the proposed innovative approach does not limit news content rating to a one-dimensional binary response. In particular, the authors utilize a vast spectrum of processing techniques, including text embedding, machine translation, and various users' preferences to provide content assessment. The results are presented using a Polish digital media use-case.

1. Introduction

1.1. Context and rationale

With the emergence of the transformer architecture, the focal point of the AI revolution shifted from computer vision to the Natural Language Processing (NLP) methods and solutions, as well as multimodal tasks. This manifests itself in the form of NLP architectures reaching sizes of hundreds of billions of parameters [1], significantly larger than those dedicated exclusively to computer vision applications. This rapid development brings obvious benefits but also intensifies existing risks and introduces new ones. One such challenge is the dissemination of fake news, propaganda, and disinformation – phenomena not new in themselves [2]. Currently, with the presence of extremely Large Language Models (LLMs), the ease and accessibility of machine-generated content has reached unprecedented levels. ChatGPT amassed 100 million subscribers within two months [3], being a tool capable of generating synthetic linguistic content indistinguishable from human-written text [4]. Such tools enable the generation of synthetic propaganda, disinformation, or fake news more easily and on a massive scale across numerous natural languages. For instance, ChatGPT supports at least 95 natural languages [5]. The widespread presence of generative NLP architectures gave rise to a new phenomenon: artificial generation of fake news via machine-generated content [6]. As a result, the threat of fake news, propaganda, and disinformation grows increasingly significant. These phenomena pose significant challenges for individuals and society at large.

Although superficially the phenomenon of fake news may seem insignificant, the COVID-19 pandemic demonstrated that the disinformation disseminated through social media can endanger many lives, as some individuals employed fictitious and toxic medical treatments [7], could lead to riots resulting in injured participants [7], influence the perception and value of companies for clients [8], or even sway election outcomes [9]. These phenomena have been observed by numerous

* Corresponding author at: Bydgoszcz University of Science and Technology, Poland.
E-mail address: chorasm@pbs.edu.pl (M. Choraś).

https://doi.org/10.1016/j.neucom.2024.128450
Received 22 March 2024; Received in revised form 27 June 2024; Accepted 19 August 2024
Available online 22 August 2024
0925-2312/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

international organizations, governments, and traditional media, evidenced by the creation of over 300 institutions dedicated to content verification and debunking of fake news [10,11]. Furthermore, the spread of fake news poses significant cybersecurity challenges, as malicious actors can exploit information systems and social media platforms to propagate disinformation, thereby undermining the integrity of digital infrastructures and eroding trust in online information sources [12].

Moreover, Generative AI, particularly when based on LLMs, has emerged as a significant threat to societies. It can be employed to craft disinformation strategies, analyzing trends, emotions, and literature to create convincing narratives or to exploit societal divisions and emotional triggers. Furthermore, LLMs facilitate the rapid generation of multiple versions of a disinformation concept, tailored for various reader groups, information bubbles, locations, and channels—be it traditional or social media. Nonetheless, there is also potential for such technology, specifically Generative AI based on LLMs, to play a crucial role in the future detection and countering of disinformation.

As pointed in [13], a multi-factor text analysis requires hybrid approach utilizing various techniques and methods that can capture different aspects of the analyzed content. In practice, tackling the complex category detection is not a trivial task to do. One of the approaches may consider

Synthetic Data Generation: Creating synthetic examples to balance classes and improve model generalization. Data Annotation: Using techniques like active learning to iteratively label data that is most informative for the model.

1.2. Paper's contributions

In this work, the very innovative results of the SWAROG project have been presented. SWAROG is an ongoing Polish research project, funded by the National Centre of Research and Development, which involves the creation of artificial intelligence algorithms for the automatic classification and detection of fake news. The focus of the project is also on building models and datasets for the Polish language, derived from real news stories in digital media.

The major contributions of this work are (i) the new and innovative proposal of the digital content assessment framework, (ii) the proposition of the innovative data processing pipeline for explainable fake news detection, (iii) the new methodology to embed annotator metrics in the feature vector together with the results of two variants of text analysis, (iv) experimental methodology to evaluate several factors relevant for fake news detection and comparison with other works.

This paper is structured as follows: Section 2 analyzes the state of the art and related work. Section 3 presents the details of the proposed approach. Section 4 provides the experiments and results. Finally, the discussion and conclusions are given in Sections 5 and 6, respectively

2. Related works

Detecting fake news and disinformation presents a challenging and demanding task, largely because natural languages embody inherent ambiguities, including sarcasm and irony. Not all texts lacking factual information constitute disinformation or fake news; some may be satirical in nature. Although various methods for detecting fake news have been developed, there is no single, universally accepted approach to this multifaceted task. The challenge is further compounded by the sheer volume and variety of texts: some are professionally produced by news agencies and seasoned journalists, while others are generated by social media users. Content from these sources varies widely, ranging from strict adherence to linguistic norms to significant deviations, including the use of specialized lingo that functions as a code within specific user groups. Another significant barrier in detecting fake news is the predominant focus on English language detection methods. Low-resource languages, including Polish, receive considerably less attention in research compared to high-resource languages. This discrepancy makes the task of automatically identifying fake news in languages other than English objectively more challenging.

Despite these challenges, advancements in the area of Natural Language Processing allow for an increased use of machine and automated detectors of fake news. In addition, there are also methods based on the analysis of sentence structures, text structures, or statistical analysis [7]. Another group of methods involves detailed source analysis and determining the reliability of the text through verification as to whether it originates from a trustworthy source with a recognized reputation [14]. There are also manual methods of fake news detection, such as engaging highly qualified journalists who analyze contents independently regarding their potential disinformation and fake news components [10,15]. The new NLP methods are used in two ways. They support the work of human fake news detectors, as some researchers argue that at this stage it is not possible to completely eliminate human annotators and specialists in identifying fake news [15]. However, there are solutions that work effectively in narrower thematic areas, such as texts related to the topic of COVID-19, or those with specific textual characteristics like social media posts (Tweets and posts). These automatic fake news detection methods have proven successful within such narrow contexts [16].

Though the linguistic methods were among the ones used the earliest for the detection of fake news, research in this area continues to emerge and yield new insights. For instance, in [7], the authors analyzed the seven features they deemed influential in identifying disinformation and fake news. These features encompassed: Sentence length, Repetition, Quotes, Word length, Technical terms, Swear words, Sexual words [7]. Conversely, in [17], the authors suggested a linguistics-based methodology to examine the lexical, syntactic, and semantic aspects of a news article for the automated identification of fake news. Additionally, [18] discusses the application of stylistic analysis on hyperpartisan (extremely one-sided) news and fake news for semi-automated detection of fake news.

The next group of methods involves manually identifying fake news by teams of journalists. Although advancements have been made in automated fact-checking systems, manual verification remains crucial in counteracting misinformation. These manual methods are described in articles [15,19]. It is assumed that NLP methods can support the work of human fake news annotators at certain stages of the process. For instance, during the competitions at CLEF2022 and CLEF2023 [20], several methods were created (based on GPT-3, mT5 or Roberta architectures) which enable detecting whether a text contains check-worthy claims, verifiable factual claims, or harmful claims. This advancement results from the insight gained into the workflows of human annotators, demonstrating that automated claim detection can effectively complement their efforts by providing valuable NLP-based support

Several studies have suggested using transformers for detecting fake news, as illustrated in [21]. In this article, models were introduced to tackle fake news classification, considering labels such as false, partially false, true, and others. Studies employed models based on the BERT (Bidirectional Encoder Representations from Transformers), XLNet, RoBERTa, and DistilBERT architectures. Similarly, in [22], researchers used BERT and developed a model named FakeBERT, which integrates multiple parallel blocks of a single-layer deep Convolutional Neural Network (CNN) with various kernel sizes and filters alongside BERT. The proposed FakeBERT model achieved an accuracy of 98.90%. Moreover, in [23], researchers built upon the BERT architecture and introduced MWPBert, which employs two parallel BERT networks for handling the headlines and bodies of news articles, respectively.

The role of extremely Large Language Models in the critical domain of countering misinformation, debunking fake news, remains largely unexplored due to the models having emerged relatively recently. Anyway, several studies have begun to include extremely LLMs in fake news detection. In [24], researchers assessed ChatGPT's fact-checking

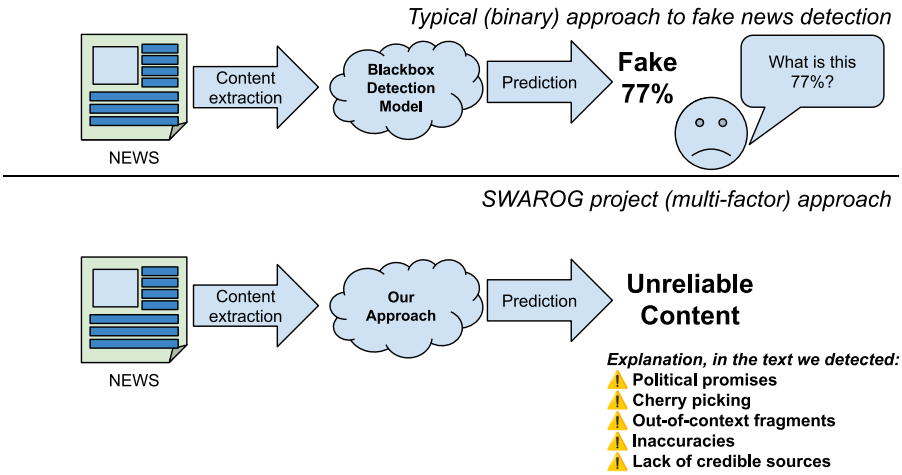


Fig. 1. The explainable news content assessment - general concept and approach proposed and used in the 'SWAROG' project. Instead of classical fake/true classification (which is obscure) the multi-factor approach has been applied, which gives the user more details about the analyzed content.

capabilities via a zero-shot classification methodology. A comprehensive comparison of four LLMs (ChatGPT3.5, Bard, BingAI, GPT4.0), conducted in [25], employed these models for News Fact-Checking tasks. Authors of [26] examine ChatGPT's capacity to identify fake news.

3. Proposed innovative approach for explainable fake news detection

3.1. The context and the news content assessment approach

Since the phenomenon of fake news is a multi-factor concept, the content rating form was not limited to a one-dimensional binary response. A set of carefully crafted questions has been developed to indicate signs of untrue content (see Fig. 1).

The work used a new Polish language dataset [13] that contains more than 40,000 annotated articles from digital newspapers along with answers to each of the 13 questions, based on the annotations that were made.

All questions took into account various aspects of the analyzed articles and were precisely formulated. They were divided into three groups: verification, manipulation, and metaphysical.

The first group of factors, verification factors, included questions that required additional activity on the part of the annotator in the form of verification of the content presented in the article in other sources.

Manipulative factors were intended to examine whether the author of the published content intended to deliberately impose a specific point of view on the reader or to mislead the recipient.

The last group of factors are metaphysical factors. The purpose of their use is to influence the reader's emotions, taking into account their views. The authors of the texts, knowing their target group, are able to influence the process of disseminating false information and give the content a specific tone.

The questions that the respondents answered during the annotation process are presented below. As it can be seen, they were precisely formulated and took into account various aspects of the analyzed articles. In the context of this research, it is important that the annotators are non-experts and they answered yes or no to the questions detailed in Table 1.

It is worth mentioning that the data set was characterized by an unbalanced distribution. There were significant discrepancies in the number of positive and negative annotations regarding individual questions (see Fig. 2).

Table 1
The list of 13 questions used during the annotation process.

No.	Question
1	Is there at least one reliable source that confirms all the information contained in the content?
2	Is most of the information provided confirmed by reliable sources?
3	Is none of the information confirmed by reliable sources?
4	Does the statement refer to current data?
5	Is additional information required to properly understand the content?
6	Does the statement contain any inaccuracies?
7	Does the statement contain fragments taken out of context?
8	Does the author of the statement use cherry-picking?
9	Is the author of the statement trying to mislead the reader?
10	Is the content satirical?
11	Does the author admit that the facts presented are made up?
12	Does the statement contain political promises?
13	Does the statement contain religious content?

3.2. The processing pipeline

The proposed approach utilizes various processing techniques, including text embedding, machine translation, annotator metric, and classification models, to classify texts in Polish and English. The key stages have been depicted in Fig. 3 and are summarized as follows:

1. The initial text in Polish or English is input into the system.
2. In the next stage, feature extraction is performed based on the transformer architecture (particularly BERT). Here, two processing variants are considered:
   - The first processing variant uses BERT Embeddings (text embeddings) for multiple languages (e.g., xlm-Roberta-multilingual).
   - The second processing variant involves an additional step of translation from Polish to English, followed by the use of BERT Embeddings for the English language.
3. Simultaneously, the annotator metric is extracted and transformed into a binary feature vector using the One-Hot Encoding technique.
4. Subsequently, feature vectors are generated based on the results of BERT Embeddings and One-Hot Encoding.
5. Feature vectors serve as input to the ensemble of N classifiers, which are classification models analyzing the content of the text regarding N questions that were answered by an expert during annotation.

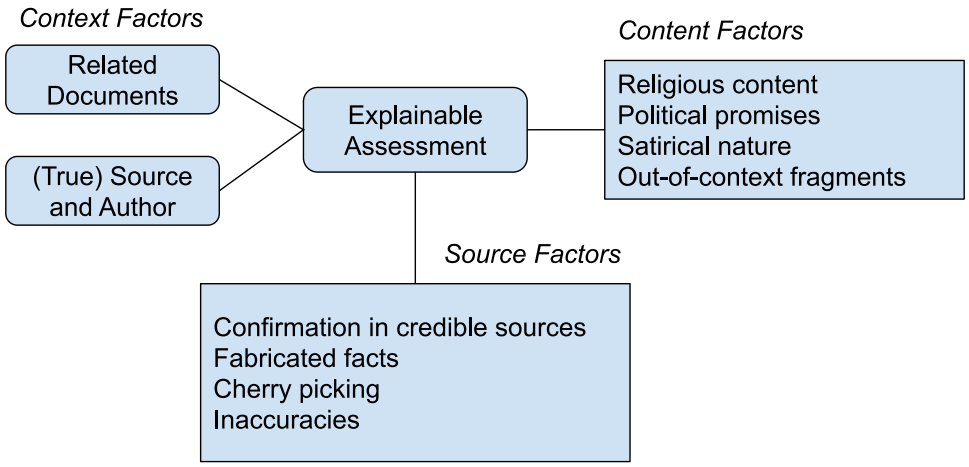


Fig. 2. The SWAROG project approach to factor-based explainable news assessment.

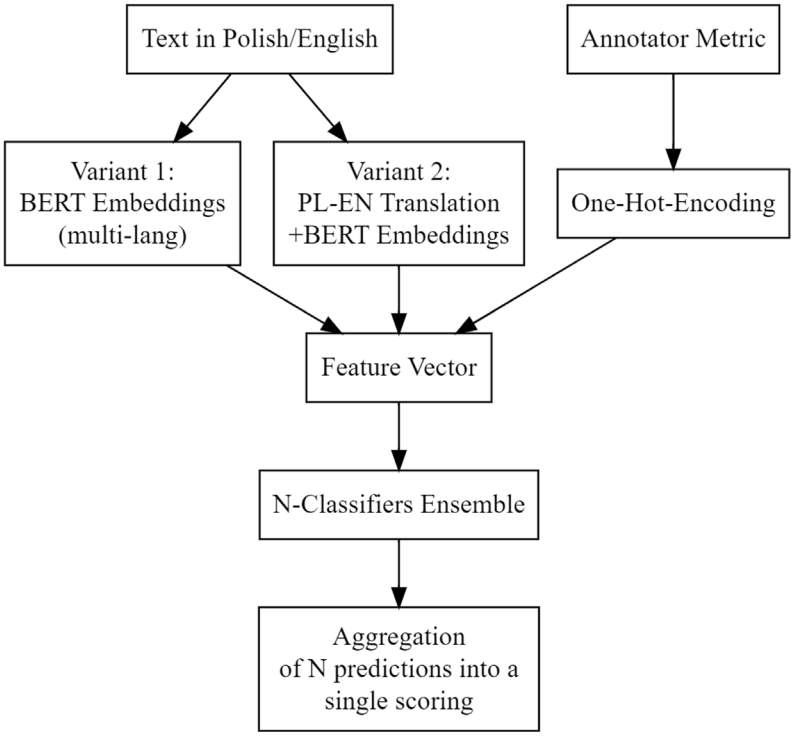


Fig. 3. General data processing pipeline used in the proposed approach.

6. To unify the system's responses for binary fake news classification, the prediction results from N classifiers are aggregated into a single scoring, which represents the final classification result.

As depicted in Fig. 3, the expert assessment metric contains categorical data, because it includes questions about gender, age, place of residence, voivodship (region), education, and profession. Moreover, the political and religious views of experts and how they use traditional and social media were also taken into account, among others. In order to handle it adequately, the One-Hot Encoding technique was used to represent this part of the data. The result of this operation is a binary vector for which each category is assigned a unique position in the vector.

It must be mentioned that in the proposed approach, various architectures based on transformer networks have been investigated. However, the one that is particularly suited to the considered context is the Sentence-BERT (S-BERT) architecture. It is an extension of the BERT model focused on sentence representation. Unlike the traditional BERT architecture, which operates at the token level, S-BERT produces feature vectors for entire sentences or text fragments. From the technical standpoint, a similar outcome can be achieved with mean/max pooling on top of the classical BERT architecture. However, the main goal of S-BERT is to generate semantic representations of sentences, which is crucial in tasks related to comparison, classification, and grouping of sentences. It must be highlighted that this process involves training the model on a large dataset of text data to capture the semantic representations of similarities between sentences.

4. Experiments and results

In this section, the experiments that had been conducted in order to assess the performance and the robustness of the proposed approach have been presented.

Table 2
Values of BAC for each fold.

Cross-validation fold	Balanced accuracy
1	0.7251
2	0.7394
3	0.7184
4	0.7265
5	0.7184
6	0.6627
7	0.7256
8	0.7125
9	0.7543
10	0.7177
Avg.	0.7207
STD.	0.0189

Table 3
BAC and standard deviation for each source.

Source	Balanced accuracy	STD
dziennikbaltycki_wiadomosci	0.60967	0.01698
dziennikpolski_wiadomosci	0.78179	0.01303
expressilustrowany_wiadomosci	0.61764	0.01844
gazetalubuska_wiadomosci	0.64883	0.01617
gazetapl_kobieta	0.77927	0.01590
gazetapl_kultura	0.82534	0.02579
gazetapl_next	0.74892	0.01834
gazetapl_plotek	0.79918	0.01980
gazetapl_podroze	0.79097	0.02398
gazetapl_sport	0.86046	0.02631
gazetapl_wiadomosci	0.75689	0.03120
gazetawroclawska_wiadomosci	0.80425	0.02494
gazetawspolczesna_wiadomosci	0.63901	0.01921
onet_wiadomosci	0.70482	0.01705
tvn24_biznes	0.79913	0.02594
tvn24_meteo	0.86371	0.02449
tvn24_sport	0.87886	0.02499
tvn24_wiadomosci	0.70824	0.02799
wp_finanse	0.77453	0.03053
wp_gwiazdy	0.81459	0.02391
wp_kobieta	0.81459	0.02391

More precisely, the authors focused on such evaluation tasks as: binary fake news classification, news source recognition, annotator's answer prediction, and comparison with other methods.

To assess the performance, common metrics were used for comparing classification effectiveness, namely:

- Precision $= \frac{TP}{TP+FP}$
- Recall (Sensitivity) $= \frac{TP}{TP+FN}$
- Specificity $= \frac{TN}{TN+FP}$
- Balanced Accuracy $= \frac{Sensitivity+Specificity}{2}$
- F1-score $= 2 * \frac{Precision*Recall}{Precision+Recall}$

One of the metrics of particular focus has been the Balanced Accuracy (BAC). It is a method of calculating model correctness when the data set is unbalanced. It represents the macro-average of recall performances in a specific class. It is important to remember that each sample is weighted based on the inverse frequency with which it occurs within a certain class.

4.1. Binary fake news classification

In this section, the multi-factor document analysis was reduced to a binary classification problem. In that regard, the responses obtained for factor predictors were aggregated in order to obtain a single score. For brevity, Table 2, presents the balanced accuracy metric (BAC) values. In order to check the robustness, the 5 × 2 fold cross validation was adopted. In the table, the global average value as well as the BAC scores for each fold have been reported.

In this experiment, the authors managed to achieve 72.1% of the BAC score. Moreover, the standard deviation from this average value was relatively low (less than 2%). In order to make this assessment more detailed, in the next section, the method has been contrasted with other state-of-the-art techniques and the obtained results were then compared on different datasets.

Table 4
Detailed classification metrics for questions Q1–Q7.

	Accuracy	BAC	F1	G-mean	Precision	Recall
Q1: At least one reliable source confirming all the information contained in the text						
Average	72.0	72.3	72.7	72.3	74.6	72.0
Yes	–	–	77.1	72.3	83.7	71.5
No	–	–	64.0	72.3	57.0	73.1
Q2: The majority of the provided information has confirmation in reliable sources						
Average	75.2	72.7	75.2	72.7	75.2	75.2
Yes	–	–	81.0	72.2	80.9	81.0
No	–	–	64.5	72.2	64.5	64.4
Q3: None of the information have confirmation in reliable source						
Average	74.0	73.8	77.4	73.8	85.2	74.0
Yes	–	–	45.5	73.8	32.9	73.6
No	–	–	82.9	73.8	94.2	74.1
Q4: The statements refer to current data						
Average	78.8	72.1	79.3	71.8	80.2	78.8
Yes	–	–	86.0	71.2	88.3	83.9
No	–	–	55.5	71.2	51.4	60.4
Q5: Necessity of Additional Information for Understanding Content						
Average	71.6	71.5	73.5	71.5	78.4	71.6
Yes	–	–	54.2	71.5	43.6	71.5
No	–	–	79.4	71.5	89.1	71.6
Q6: Statement Contains Inaccuracies						
Average	71.0	71.2	73.6	71.2	80.1	71.0
Yes	–	–	50.0	71.2	38.5	71.4
No	–	–	79.6	71.2	90.7	70.9
Q7: Presence of Out-of-Context Fragments						
Average	73.1	72.7	76.3	72.7	83.5	73.1
Yes	–	–	46.7	72.7	34.5	72.0
No	–	–	82.0	72.7	93.1	73.3

4.2. News source recognition

Another aspect of the multi-factor analysis is the possibility of true source recognition. In that regard, by having the text of a news article, the authors try to predict its true origins using the list of known sources. Similarly, as in the previous case, the 5 × 2 cross-validation technique was adopted. Therefore, in Table 3, for each source, the average BAC scores along with the standard deviation from the mean values were presented.

4.3. Question prediction

This section presents the results obtained for predicting individual annotator responses using the proposed model. More precisely, each question is considered a binary classification problem, as the expert was asked to answer with either 'yes' or 'no' for each question. In tables Tables 4 and 5, the values for various classification metrics were presented.

In general, the classification performance ranges from 70% to 90% of accuracy. This outcome allows one to hypothesize that some questions are difficult to provide single and objective answers because

Table 5
Detailed classification metrics for questions Q8–Q13.

	Accuracy	BAC	F1	G-mean	Precision	Recall
Q8: Use of Cherry Picking						
Average	79.3	72.2	80.8	71.9	83.4	79.3
Yes	–	–	49.4	71.4	41.2	61.7
No	–	–	87.0	71.4	91.7	82.8
Q9: Attempt to Mislead Readers						
Average	92.1	79.4	93.3	78.4	95.3	92.1
Yes	–	–	42.7	78.2	31.6	65.5
No	–	–	95.7	78.2	98.3	93.3
Q10: Presence of Satirical Content						
Average	95.2	79.8	96.4	78.2	98.1	95.2
Yes	–	–	29.8	78.1	19.4	63.8
No	–	–	97.5	78.1	99.4	95.7
Q11: Acknowledgment of Fabricated Facts						
Average	95.6	82.4	97.1	81.3	99.1	95.6
Yes	–	–	19.6	81.3	11.4	69.0
No	–	–	97.7	81.3	99.7	95.8
Q12: Presence of Political Promises						
Average	91.2	90.8	93.0	90.8	96.3	91.2
Yes	–	–	50.1	90.8	34.6	90.4
No	–	–	95.2	90.8	99.5	91.3
Q13: Presence of Religious Content						
Average	95.7	81.9	96.9	80.7	98.6	95.7
Yes	–	–	26.9	80.7	16.8	67.8
No	–	–	97.8	80.7	99.6	96.0

background knowledge might be required. For example, the detection of cherry-picking requires a good understanding of a particular topic, which often spans beyond the original text. Therefore, the accuracy of detecting that kind of matter is relatively lower when compared to less complex questions, like those asking whenever the content is to be considered religious. In fact, judging news for religious or political content does not require additional context (beyond this contained in the original text). Therefore, in this case, the classifiers achieve higher accuracy (e.g. from 80% to 90% for the BAC metric).

4.4. Comparison to other approaches

In order to compare the proposed approach with other approaches, the PubHealth dataset [27] was used. The dataset contains documents related to public health, including biomedical subjects (e.g., infectious diseases, stem cell research) and government healthcare policy (e.g., abortion, mental health, and women's health). Each instance in the dataset has an associated label (true, false, unproven, mixture). Moreover, each instance in the dataset has an explanation text field. The explanation is a justification for which the claim has been assigned a particular label.

The considered dataset contains original news texts which are annotated using 4 types of labels, namely: False News, True News, Mixture, and Unproven. Moreover, the creators have split the entire collection into training and testing subsets. Similarly, as the competing methods, the training dataset has been used to fine-tune the solutions presented. The results from the experiments have been shown in Table 6. In general, the achieved classification performance is comparable to the other considered methods. However, in this case, thanks to the multi-factor-based approach, a layer of self-explainability has also been provided, which spans beyond typical word-based attribution.

5. Discussion and threats to validity

The results presented in this work are based on the dataset created in the SWAROG project. Through a paid service, external annotators were engaged to answer several questions about the content. This method provided the authors with a richer dataset for analysis beyond mere binary classification. Of course, in such an approach and project, it means that the results depend on the annotators and the wide set of digital content that was analyzed.

Table 6
Comparison on PubHealth dataset (accuracy scores).

Method	False	Mixture	True	Unproven
BB-Health [28]	0.809	0.498	0.915	0.4
Longformer-Health2 [29]	0.809	0.532	0.908	0.4
Proposed approach	0.789	0.10	0.87	0.2

The authors acknowledge that their methodology might yield slightly different results when applied to other languages, content types, or with annotations from different individuals. Despite these potential variations, the dataset presented is considerably rich. The approach discussed, detailed in Section 3, incorporates metrics about the annotators themselves (such as region, opinions, beliefs, etc.). This strategy makes it possible to visually represent the outcomes from various annotator subsets. It not only enhances the explainability of the findings but also enables the analysis of results within specific 'information bubbles'.

In the approach presented, information bubbles and multi-factor results contribute to the explainability of the fake news detection system (together with xAI methods, as e.g. presented in [30]). However, the authors are aware that explainability in fake news detection systems can also be used for adversarial attacks, as they earlier showed and proved in [31]. Consequently, the application of explainable approaches must be approached with caution to ensure the system remains robust against manipulation.

6. Conclusions

Transformer architectures are now being widely used to combat the phenomenon of fake news. Unfortunately, a significant number of solutions have started to treat misinformation in a similar way as other natural language processing challenges (e.g., sentiment classification). The research presented in this paper contradicts this by approaching the phenomenon of fake news from a multi-factor perspective. The proposed approach is not limited to a one-dimensional binary news content rating. It aims at embracing the reader's preferences along with the text content analysis. In order to evaluate the proposed solution, the multi-factor dataset released by the SWAROG project was used. The authors have also compared their method with other techniques available in the literature. The method presented not only enables one to achieve results that are comparable with existing approaches but also enhances the analysis by incorporating a factor-based assessment.

CRediT authorship contribution statement

Rafał Kozik: Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Gracjan Kątek:** Writing – original draft, Validation, Software, Investigation. **Marta Gackowska:** Investigation, Data curation. **Sebastian Kula:** Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Joanna Komorniczak:** Software, Methodology, Formal analysis, Data curation, Conceptualization. **Paweł Ksieniewicz:** Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Aleksandra Pawlicka:** Writing – original draft, Validation. **Marek Pawlicki:** Writing – original draft, Formal analysis. **Michał Choraś:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rafał Kozik reports financial support was provided by the National center for Research and Development within INFOSTRATEG program and by Horizon Europe. Gracjan Katek reports financial support was provided by the National center for Research and Development within INFOSTRATEG program. Marta Gackowska reports financial support was provided by the National center for Research and Development within INFOSTRATEG program. Sebastan Kula reports financial support was provided by the National center for Research and Development within INFOSTRATEG program. Joanna Komorniczak reports financial support was provided by the National center for Research and Development within INFOSTRATEG program. Paweł Ksieniewicz reports financial support was provided by the National center for Research and Development within INFOSTRATEG program. Marek Pawlicki reports financial support was provided by the National center for Research and Development within INFOSTRATEG program and by Horizon Europe. Aleksandra Pawlicka reports financial support was provided by Horizon Europe. Michał Choraś reports financial support was provided by the National center for Research and Development within INFOSTRATEG program and by Horizon Europe.

Data availability

Data will be made available on request.

Acknowledgments

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.

This work is partially funded under the HE AI4CYBER project, with the support of the European Commission and the Horizon Europe Programme, under Grant Agreement No. 101070450.

References

[1] S. Krishna, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, H. Lakkaraju, Post hoc explanations of language models can improve language models, 2023, http://dx.doi.org/10.48550/ARXIV.2305.11426, CoRR abs/2305.11426 arXiv:2305.11426.

[2] S. Kula, R. Kozik, M. Choras, M. Wozniak, Transformer based models in fake news detection, in: M. Paszynski, D. Kranzlmüller, V.V. Krzhizhanovskaya, J.J. Dongarra, P.M.A. Sloot (Eds.), Computational Science - ICCS 2021 - 21st International Conference, Krakow, Poland, June 16-18, 2021, Proceedings, Part IV, in: Lecture Notes in Computer Science, vol. 12745, Springer, 2021, pp. 28–38, http://dx.doi.org/10.1007/978-3-030-77970-2_3.

[3] B. Niu, G.F.N. Mvondo, I am ChatGPT, the ultimate AI Chatbot! investigating the determinants of users' loyalty and ethical usage concerns of ChatGPT, J. Retail. Consum. Serv. 76 (2024) 103562, http://dx.doi.org/10.1016/j.jretconser.2023.103562, URL https://www.sciencedirect.com/science/article/pii/S0969698923003132.

[4] S. Mitrović, D. Andreoletti, O. Ayoub, ChatGPT or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text, 2023, arXiv:2301.13852.

[5] M. Alawida, S. Mejri, A. Mehmood, B. Chikhaoui, O. Isaac Abiodun, A comprehensive study of ChatGPT: Advancements, limitations, and ethical considerations in natural language processing and cybersecurity, Information 14 (8) (2023) http://dx.doi.org/10.3390/info14080462, URL https://www.mdpi.com/2078-2489/14/8/462.

[6] K. Shu, Y. Li, K. Ding, H. Liu, Fact-enhanced synthetic news generation, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press, 2021, pp. 13825–13833, http://dx.doi.org/10.1609/AAAI.V35I15.17629.

[7] A. Krickl, S. Kirrane, Misinformation detection: Using linguistic cues, in: U. Simsek, D. Chaves-Fraga, T. Pellegrini, S. Vahdat (Eds.), Proceedings of Poster and Demo Track and Workshop Track of the 18th International Conference on Semantic Systems Co-Located with 18th International Conference on Semantic Systems, SEMANTiCS 2022, Vienna, Austria, September 13th to 15th, 2022, in: CEUR Workshop Proceedings, vol. 3235, CEUR-WS.org, 2022, URL https://ceur-ws.org/Vol-3235/paper9.pdf.

[8] M.C. Arcuri, G. Gandolfi, I. Russo, Does fake news impact stock returns? Evidence from US and EU stock markets, Journal of Economics and Business 125–126 (2023) 106130, http://dx.doi.org/10.1016/j.jeconbus.2023.106130, URL https://www.sciencedirect.com/science/article/pii/S0148619523000231.

[9] N. Hassan, F. Arslan, C. Li, M. Tremayne, Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017, ACM, 2017, pp. 1803–1812, http://dx.doi.org/10.1145/3097983.3098131.

[10] E. Pankovska, K. Schulz, G. Rehm, Suspicious sentence detection and claim verification in the COVID-19 domain, in: ROMCIR@ECIR, 2022, URL https://api.semanticscholar.org/CorpusID:249283424.

[11] E.R. Mark Stencel, J. Luther, Fact-checkers extend their global reach with 391 outlets, but growth has slowed, 2024, https://www.poynter.org/fact-checking/2022/391-global-fact-checking-outlets-slow-growth-2022/. (Accessed 5 January 2024).

[12] M. Choras, A. Pawlicka, M. Pawlicki, R. Kozik, How to navigate the uncharted waters of cyberwar, in: AMCIS 2023 Proceedings, 2023, URL https://aisel.aisnet.org/amcis2023/conf_theme/conf_theme/1/.

[13] 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.

[14] N. Seddari, A. Derhab, M. Belaoued, W. Halboob, J. Al-Muhtadi, A. Bouras, A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media, IEEE Access PP (2022) 1, URL https://api.semanticscholar.org/CorpusID:249526926.

[15] R. Procter, M. Arana-Catania, Y. He, M. Liakata, A. Zubiaga, E. Kochkina, R. Zhao, Some observations on fact-checking work with implications for computational support, 2023, http://dx.doi.org/10.48550/ARXIV.2305.02224, CoRR abs/2305.02224 arXiv:2305.02224.

[16] R.A. Dar, R. Hashmy, A survey on COVID-19 related fake news detection using machine learning models, in: MoMLeT+DS, 2023, URL https://api.semanticscholar.org/CorpusID:259254278.

[17] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: E.M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, Association for Computational Linguistics, 2018, pp. 3391–3401, URL https://aclanthology.org/C18-1287/.

[18] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A stylometric inquiry into hyperpartisan and fake news, 2017, arXiv arXiv:1702.05638 URL https://api.semanticscholar.org/CorpusID:574574.

[19] A. Hrckova, R. Móro, I. Srba, J. Simko, M. Bieliková, Automated, not automatic: Needs and practices in European fact-checking organizations as a basis for designing human-centered AI systems, 2022, http://dx.doi.org/10.48550/ARXIV.2211.12143, CoRR abs/2211.12143 arXiv:2211.12143.

[20] P. Nakov, A. Barrón-Cedeño, G.D.S. Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y.S. Kartal, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, in: CEUR Workshop Proceedings, vol. 3180, CEUR-WS.org, 2022, pp. 368–392, URL http://ceur-ws.org/Vol-3180/paper-28.pdf.

[21] S. Althabiti, M.A. Alsalka, E. Atwell, SCUoL at CheckThat!-2022: Fake news detection using transformer-based models, in: Conference and Labs of the Evaluation Forum, 2022, URL https://api.semanticscholar.org/CorpusID:251471040.

[22] R.K. Kaliyar, A. Goswami, P. Narang, FakeBERT: Fake news detection in social media with a BERT-based deep learning approach, Multimedia Tools Appl. 80 (8) (2021) 11765–11788, http://dx.doi.org/10.1007/s11042-020-10183-2.

[23] M. Farokhian, V. Rafe, H. Veisi, Fake news detection using dual BERT deep neural networks, Multimedia Tools Appl. (2022) URL https://api.semanticscholar.org/CorpusID:248085291.

[24] E. Hoes, S. Altay, J. Bermeo, Leveraging ChatGPT for efficient fact-checking, 2023, http://dx.doi.org/10.31234/osf.io/qnjkf, PsyArXiv.

[25] K.M. Caramancion, News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, bing AI, and bard in news fact-checking, 2023, http://dx.doi.org/10.48550/arXiv.2306.17176, CoRR abs/2306.17176 arXiv:2306.17176.

[26] Y. Huang, L. Sun, Harnessing the power of ChatGPT in fake news: An in-depth exploration in generation, detection and explanation, 2023, arXiv arXiv:2310.05046 URL https://api.semanticscholar.org/CorpusID:263831595.

[27] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 7740–7754, URL https://www.aclweb.org/anthology/2020.emnlp-main.623.

[28] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big bird: Transformers for longer sequences, 2021, arXiv:2007.14062.

[29] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The long-document transformer, 2020, arXiv:2004.05150.

[30] M. Szczepański, M. Pawlicki, R. Kozik, M. Choraś, New explainability method for BERT-based model in fake news detection, Sci. Rep. 11 (2021) 23705, http://dx.doi.org/10.1038/s41598-021-03100-6, URL https://www.nature.com/articles/s41598-021-03100-6.

[31] 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 (2024) 103599, http://dx.doi.org/10.1016/j.cose.2023.103599, URL https://www.sciencedirect.com/science/article/pii/S0167404823005096.