big data and cognitive computing

MDPI

Article

Exploiting Content Characteristics for Explainable Detection of Fake News

Sergio Muñoz * **and Carlos Á. Iglesias**

Intelligent Systems Group, Telematic Systems Engineering Department, Universidad Politécnica de Madrid, Avenida Complutense 30, 28040 Madrid, Spain; carlosangel.iglesias@upm.es

* Correspondence: sergio.munoz@upm.es

Abstract: The proliferation of fake news threatens the integrity of information ecosystems, creating a pressing need for effective and interpretable detection mechanisms. Recent advances in machine learning, particularly with transformer-based models, offer promising solutions due to their superior ability to analyze complex language patterns. However, the practical implementation of these solutions often presents challenges due to their high computational costs and limited interpretability. In this work, we explore using content-based features to enhance the explainability and effectiveness of fake news detection. We propose a comprehensive feature framework encompassing characteristics related to linguistic, affective, cognitive, social, and contextual processes. This framework is evaluated across several public English datasets to identify key differences between fake and legitimate news. We assess the detection performance of these features using various traditional classifiers, including single and ensemble methods and analyze how feature reduction affects classifier performance. Our results show that, while traditional classifiers may not fully match transformer-based models, they achieve competitive results with significantly lower computational requirements. We also provide an interpretability analysis highlighting the most influential features in classification decisions. This study demonstrates the potential of interpretable features to build efficient, explainable, and accessible fake news detection systems.

Keywords: fake news detection; explainability; machine learning; text classification

check for updates

Citation: Muñoz, S.; Iglesias, C.Á. Exploiting Content Characteristics for Explainable Detection of Fake News. *Big Data Cogn. Comput.* **2024**, *8*, 129. https://doi.org/10.3390/bdcc8100129

Academic Editor: Alberto Abelló

Received: 12 August 2024
Revised: 24 September 2024
Accepted: 1 October 2024
Published: 4 October 2024

1. Introduction

The digital age has revolutionized the way information is disseminated and consumed. While this transformation has brought numerous benefits, it has also led to the proliferation of fake news [1], which poses a significant threat to the integrity of information ecosystems [2]. Fake news can be described as a form of online disinformation that intentionally contains false or misleading information [3,4]. This type of content mimics the structure of legitimate news to capture the audience's attention. By intentionally deceiving or manipulating people's emotions and beliefs, fake news can influence public opinion, exacerbate social polarization, and undermine trust in legitimate news sources [5].

The implications of fake news have far-reaching economic, social, and political consequences [6]. Economically, fake news can disrupt markets, impact stock prices, and tarnish the reputations of businesses and individuals. An example of this occurred in 2013 when a false report of explosions at the White House led to a sudden drop in the S&P 500 index [7]. Socially, fake news undermines public trust in critical institutions, including the media, government, and science. It can propagate misinformation about essential health issues, as evidenced during the COVID-19 pandemic, with the spread of false information about vaccines and treatments [8]. Politically, fake news can influence elections, manipulate public opinion, and destabilize democratic processes. For example, fake news stories propagated through social media to influence voter behavior have been observed during presidential elections in several countries [9–12].

Therefore, addressing the fake news problem is crucial to preserving the integrity of information ecosystems, maintaining economic stability, fostering social cohesion, and safeguarding democratic institutions [13].

Recent advances in artificial intelligence have shown significant promise in tackling this challenge [14]. Early approaches relied on hand-crafted features like word frequency, sentiment and syntax patterns [15,16]. Additionally, word embeddings enhanced models by representing words in dense, continuous vector spaces that capture semantic relationships [17]. The introduction of deep learning and neural networks further improved performance by enabling models to learn more complex patterns in text [18]. More recently, transformer-based models have revolutionized the field using self-attention mechanisms to understand intricate language patterns and dependencies [19]. These models have set new benchmarks in accuracy and robustness for various natural language processing tasks, including fake news detection [20,21]. Despite their effectiveness, transformer-based models pose significant challenges that hinder their widespread adoption in practical fake news detection applications [22]. The high computational costs associated with these models make them resource-intensive, limiting their deployment in scenarios where real-time processing or large-scale analysis is required. Moreover, the lack of interpretability in these complex models presents a critical issue, especially in the context of fake news detection. Understanding the reasoning behind a classification is crucial in this context for building trust in the system and providing actionable insights to users and fact-checkers.

In light of these challenges, we propose an approach to strike a balance between performance, computational efficiency, and model transparency. To this aim, we propose a comprehensive set of features to determine whether differences in content characteristics can foster interpretable distinctions between fake and legitimate news. Specifically, we seek to address the following research questions:

- **RQ1:** *How do content characteristics differentiate between fake and legitimate news?* While previous studies have explored various features for fake news detection, a comprehensive analysis of how diverse content characteristics differ between fake and legitimate news is still lacking. This question aims to fill this gap by thoroughly examining linguistic, moral, affective, perceptual, social, and cognitive features. Understanding these differences is crucial for developing more nuanced and accurate detection methods.
- **RQ2:** *To what extent can traditional classifiers achieve competitive performance in fake news detection compared to advanced transformer-based models?* Recent research has focused on complex deep-learning models, particularly transformer-based architectures. However, these models often require significant computational resources and lack interpretability. By comparing traditional classifiers to state-of-the-art models, we address the critical need for efficient and transparent solutions that can be readily deployed in real-world applications.
- **RQ3:** *What is the impact of feature reduction on the effectiveness and efficiency of fake news detection systems?* The trade-off between model complexity and performance is a persistent challenge in machine learning. This question addresses the gap in understanding how feature reduction affects the performance of fake news detection systems. By exploring this relationship, we aim to contribute to more practical and scalable solutions for real-world deployment.

To address these research questions, this study introduces a comprehensive feature framework that includes attributes of various types, such as linguistic, moral, and affective values and perceptual, social, and cognitive processes. By examining how these features differ between fake and legitimate news, we aim to identify key indicators to improve the explainability and effectiveness of fake news detection systems. We assess the detection performance of these features using traditional classifiers, including single and ensemble algorithms. To this aim, we use a diverse set of public English datasets and compare the results of our approach to state-of-the-art transformer-based models. Additionally, we investigate how reducing the number of features impacts the performance and efficiency of our approach, seeking to balance accuracy with computational feasibility. Finally, we analyze the explainability of our method to provide insights into how different features contribute to classification decisions and enhance the transparency of the detection process.

The remainder of this article is organized as follows: Section 2 reviews related work in fake news detection, outlining previous approaches and highlighting key advancements and challenges. Section 3 details the methodology employed in this study, including the design of the feature framework, the selection of datasets, and the development of the fake news detection model. Section 4 presents the evaluation results, comparing our approach to transformer-based models and exploring the effects of feature reduction. It also includes a detailed analysis of our approach's explainability, focusing on how different features influence classification decisions and enhance system interpretability. Finally, Section 5 concludes the paper by summarizing the findings, discussing the implications of our results, and suggesting directions for future research.

2. Related Work

Fake news consists of verifiably false or misleading information intentionally presented as news [3,4]. This phenomenon emerges with the intent to deceive readers for various purposes, such as political gain or financial profit [23]. It comprises several types of content, such as fabricated, manipulated, and misleading content, each varying in its degree of falsehood and intent. The pervasive nature of fake news has prompted extensive research into effective detection methodologies, driven by the need to maintain public trust and safeguard the integrity of information ecosystems [24].

Detection methods for fake news can broadly be categorized into content-based and social context-based approaches [25]. Content-based methods analyze the news articles' features, such as text, images, and videos, looking for deceptive patterns or anomalies. These methods can be further divided into textual features [26], style-based features [27], and visual-based features [28]. In addition to analyzing the content, social context-based approaches examine the dissemination patterns and user interactions on social media platforms [29]. Features such as user profiles, network connections, and propagation dynamics enhance detection accuracy. For example, Shu et al. [29] showed how user interactions and publisher–news relations can complement content features to improve detection accuracy. By leveraging social network analysis and graph-based methods, researchers can better understand how fake news spreads and identify key indicators of false information.

Still, the ubiquitous availability of textual content in news articles has made textual analysis one of the most prominent solutions for fake news detection. In this context, using advanced natural language processing techniques has enabled the development of robust and accurate detection models. Early efforts in fake news detection employed traditional machine learning classifiers trained with manually crafted feature sets, including lexical and syntactic attributes [30]. While these methods provided a foundation for fake news detection, they often struggled with capturing the nuanced semantics of language and context. The introduction of word embeddings, such as Word2Vec, GloVe, and FastText, marked a significant advancement. This technology enables the representation of words as dense, continuous vector spaces that capture semantic relationships between them. Word embeddings allow models to understand context and meaning more effectively, improving performance [31]. Combining traditional classifiers with word embeddings has proven a powerful approach in different text classification tasks [17,32,33]. However, these methods still face limitations in capturing long-range dependencies and contextual information across sentences and paragraphs.

The rapid evolution of deep learning has led to more sophisticated models [18,34]. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) improved upon traditional models by capturing sequential and contextual information in text [35,36]. These models can also process and integrate various forms of data, enabling the development of multimodal detection systems that analyze text, images, and videos simultaneously [37,38].

More recently, the introduction of transformer-based models has revolutionized natural language processing tasks [39], including fake news detection. These models, pre-trained on large datasets, use self-attention mechanisms to understand intricate language patterns and dependencies, setting new benchmarks in accuracy and robustness. Fine-tuning these models on domain-specific data further boosts their performance [40,41]. Despite their impressive performance, transformer-based models often require substantial computational resources [42], making them expensive to train and deploy. Additionally, their complex architectures and large parameter spaces contribute to a lack of interpretability [22], making it challenging to understand how they arrive at specific decisions. This complexity can hinder the practical implementation of these models in real-world applications where transparency and resource efficiency are crucial. Understanding how models arrive at their decisions is crucial for building trust and ensuring ethical AI deployment [43]. For this reason, recent studies have integrated explainability techniques, such as SHAP (Shapley additive explanations) or LIME (local interpretable model-agnostic explanations) [44], to provide insights into model predictions and identify the most influential features in classification decisions [45].

Recent advancements in fake news detection have focused on improving accuracy, robustness, and interpretability by integrating diverse information sources and methodologies. Zhang et al. [46] introduced a logic-based model for multimodal misinformation detection, integrating interpretable logic clauses to express the reasoning process. Similarly, Han et al. [47] proposed a method combining a dual graph neural network (DGNN) to optimize graph structures by eliminating redundant edges and a novel explainable reasoning module (ERM) to highlight critical nodes that support the classification. These recent studies highlight the ongoing evolution of fake news detection methodologies, emphasizing the importance of interpretability. Although significant progress has been made, challenges remain in balancing model complexity, performance, and practical applicability.

Our work aims to advance the explainable detection of fake news by introducing a comprehensive feature framework that integrates various features based on content characteristics. Through the integration of these diverse features with traditional classifiers, we aim to achieve competitive performance while maintaining transparency and practical usability. By focusing on interpretable models and a rich feature set, our approach seeks to address the limitations of complex deep-learning models while benefiting from recent advancements in the field.

3. Methodology

This section details the methodological approach adopted to address the problem of explainable fake news detection. Our approach is divided into two main components: the feature framework and the machine learning model. The feature framework includes a wide range of attributes that capture different aspects of news content, while the machine learning model evaluates the effectiveness of these features in detecting fake news.

The first step of our methodology is constructing a feature framework that captures the differences between fake and legitimate news. Based on previous studies and linguistic and psychological theory [16,48–50], we have identified five main categories of characteristics, each of which covers a specific set of relevant attributes. These five categories—linguistic, affective, cognitive, social, and contextual—have been chosen because they cover the critical aspects of text analysis that have been empirically linked to the distinguishing characteristics of fake news [51,52]. Figure 1 shows an overview of the proposed framework.

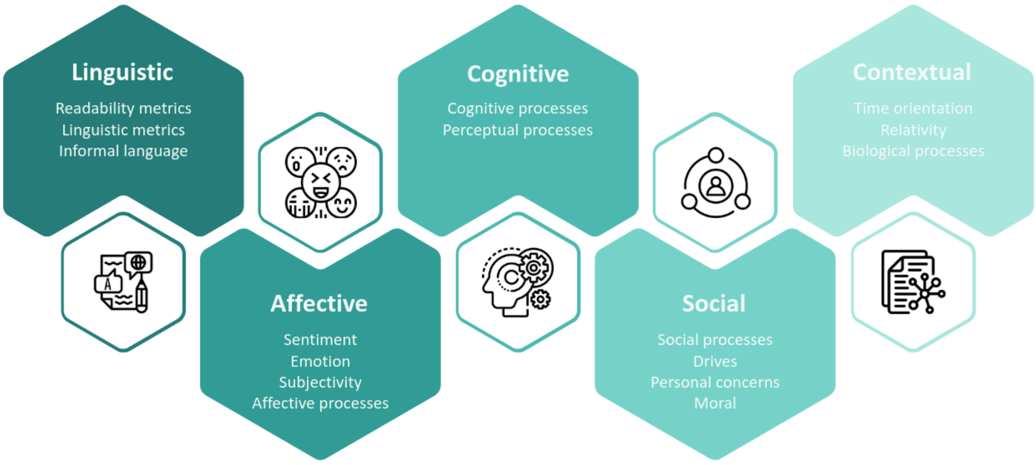

Figure 1. Comprehensive feature framework for fake news detection.

Linguistic features capture the structural and readability aspects of the text. This category includes readability grades, sentence info, and word usage (e.g., `words_per_sentence`, `complex_words`) to identify patterns in narrative style. It also includes additional insights related to word counts, usage patterns, and informal language extracted from LIWC lexicon [53] (e.g., `liwc_auxverb`, `liwc_netspeak`). These features help assess the complexity and clarity of the text, which are crucial in distinguishing between the often simplistic or convoluted language of fake news and the typically more balanced language of legitimate news.

Affective features capture the affective tone and subjective aspects of the text. These include sentiment and emotion scores derived from VADER and NRC lexicon, respectively (e.g., `sentiment_pos`, `emotion_anger`), the degree of subjectivity, and insights related to affective processes extracted from the LIWC lexicon (e.g., `liwc_posemo`, `liwc_affect`). These features are essential because fake news often leverages strong emotional appeals and subjective content to manipulate readers.

Cognitive features relate to mental processes and perceptions. This category comprises features related to cognitive and perceptual processes extracted from the LIWC lexicon (e.g., `liwc_hear`, `liwc_cause`, `liwc_see`). These features are essential for understanding the depth of content and the writer's intention, which helps identify the often shallow or biased cognitive processes employed in fake news.

Social features reflect social interactions and personal concerns. This category includes features extracted from the LIWC lexicon related to social interactions (e.g., `liwc_family`, `liwc_friend`), personal concerns (e.g., `liwc_home`, `liwc_money`), or motivational aspects (e.g., `liwc_power`, `liwc_risk`) and moral features derived from the MoralStrength lexicon (e.g., `moral_care`, `moral_fairness`) [54]. These features reflect social interactions, beliefs, and personal interests in the text, which may be exploited in fake news to resonate with or incite particular social groups.

Contextual features provide context regarding time, relativity, biological processes, and moral perspectives. Features related to time, relativity, or biological processes are extracted from the LIWC lexicon (e.g., `liwc_focuspast`, `liwc_motion`, `liwc_health`). By examining these contextual elements, we can gain insights into how fake news articles position their narratives within broader social and temporal frameworks to achieve their deceptive aims.

The readability metrics used in this study were computed with the Python readability library (Available at https://pypi.org/project/readability/ (accessed on 15 June 2024)), which provides standardized measures like Flesch–Kincaid and ARI. The TextBlob library (Available at https://pypi.org/project/textblob (accessed on 15 June 2024)) was employed to extract subjectivity scores, while MoralStrength (Available at https://pypi.org/project/moralstrength/, accessed on 15 June 2024) was used to compute moral features by analyzing text based on moral foundations theory. The liwc package (Available at https://pypi.org/project/liwc/, accessed on 15 June 2024) was used to extract a variety of linguistic, cognitive, and emotional features from the text, based on the LIWC 2015 lexicon. Finally, we utilized vaderSentiment (Available at https://pypi.org/project/vaderSentiment/, accessed on 15 June 2024) for VADER-based sentiment analysis and NRCLex (Available at https://pypi.org/project/NRCLex/, accessed on 15 June 2024) for extracting emotion features from the NRC lexicon.

As a result, a comprehensive set of more than 100 distinct features encompassing various types and characteristics is obtained. After defining the feature framework, the next step is to exploit it through a machine learning model to detect fake news. The challenge at hand is to analyze the content of a news article (n_i) to elucidate whether it is fake or legitimate news. To achieve this, we propose a comprehensive machine learning model that exploits the feature framework described above. An overview of the proposed approach can be seen in Figure 2.

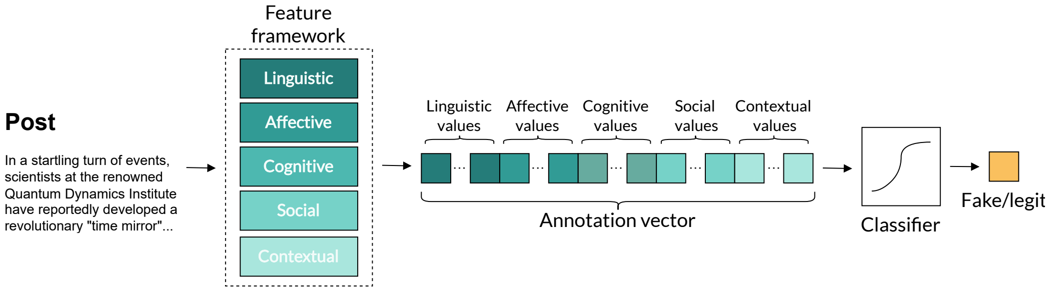

Figure 2. Proposed fake news detection model.

Specifically, our model integrates the feature framework discussed above. The feature vector for each news instance (n_i) is constructed by extracting features from the categories above. Given the set of categories considered, $C = \{C_L, C_A, C_{Cg}, C_S, C_C\}$, where C_L, C_A, C_{Cg}, C_S, and C_C correspond to linguistic, affective, cognitive, social, and contextual features, respectively, we define the set of features included in our framework as follows:

$$F \equiv \{f_{c,n} \mid c \in C, n \in [1, N_c]\} = F_L \cup F_A \cup F_{Cg} \cup F_S \cup F_C$$

where F_L, F_A, F_{Cg}, F_S, and F_C are the sets of features, and N_c is the number of features included in each category. Given a new instance n_i and the complete set of features F, an annotation vector A_i is generated. This vector contains numeric values representing the intensity of each feature in the instance:

$$A_i = \{a_{f_{c,n}} \mid f_{c,n} \in F\}.$$

Algorithm 1 illustrates the step-by-step construction of the feature vector A_i. For each category c within the framework, the corresponding set of features F_c is retrieved. Then, the annotation function is used to determine the value of each feature $f_{c,n}$ for the instance n_i. The resulting annotations $a_{f_{c,n}}$ are concatenated using the `concat` function to form the final annotation vector A_i, which encapsulates the linguistic patterns, emotional tones, cognitive signals, social interactions, and contextual elements present in the news instance.

In this way, for each news instance n_i, a comprehensive feature vector A_i is computed. Subsequently, A_i is fed into the machine learning classifier, which processes these features and returns a prediction about whether the news instance is fake or legitimate.

Algorithm 1 Construction of Feature Vector A_i

```
Input: News instance $n_i$, Feature categories $C = \{C_L, C_A, C_{Cg}, C_S, C_C\}$
Output: Annotation vector $A_i$
$A_i \leftarrow []$
for each category $c$ in $C$ do
    $F_c \leftarrow$ extract features for category $c$
    for each feature $f_{c,n}$ in $F_c$ do
        $a_{f_{c,n}} \leftarrow$ annotate $n_i$ using feature $f_{c,n}$
        $A_i \leftarrow \texttt{concat}(A_i, a_{f_{c,n}})$
    end for
end for
return $A_i$
```

4. Evaluation

To assess the effectiveness of the proposed fake news detection framework, we conducted a comprehensive evaluation consisting of several key components. This section outlines the datasets used for analysis, presents a preliminary examination of the features, details the performance results of our approach, and explores the explainability of our detection system.

4.1. Datasets

We utilized several public English datasets widely used in fake news detection research to evaluate the proposed framework. These datasets were chosen due to their diversity in terms of news sources, topics, and contexts. This diversity contributes to a comprehensive assessment of the model's performance. Specifically, the datasets vary in size, genre (e.g., political, celebrity, satirical), and linguistic complexity, providing a robust environment for evaluating the generalization of the model across different types of fake news. The primary datasets include the ISOT dataset [55], FakeNewsNet [56], FakeNewsKaggle [57], FakeNewsAMT [58], FakeNewsCelebrity [58], FakeNewsBuzFeedPolitical [59], FakenEwsRandomPolitical [59], FakeNewsPolitFalse [60], and FakeNewsSatirical [61].

Table 1 shows the statistical characteristics of the nine datasets used in this study. It presents their size (total number of posts), distribution between fake and legitimate news articles, and text length metrics (average word count and character count). The datasets vary significantly in size, ranging from 101 posts (FakeNewsBuzfeedPolitical) to 43,729 posts (ISOT). Most datasets maintain a relatively balanced distribution between fake and legitimate news, with some exceptions like FakeNewsKaggle, which has a slight imbalance favoring legitimate news. Text length also varies considerably across datasets, with average word counts ranging from 123 (FakeNewsAMT) to 936 (FakeNewsBuzfeedPolitical) and average character counts from 735 (FakeNewsAMT) to 5572 (FakeNewsBuzfeedPolitical). This diversity in dataset size, balance, and text length characteristics provides a comprehensive basis for analyzing fake news detection across different contexts and content types. The same pre-processing procedures were employed in all datasets: expansion of contractions (e.g., it's, we'll), conversion of chat words (e.g., AFAIK, ASAP), and spelling checking.

Table 1. Statistics of the used datasets.

Dataset	**No. Posts (Fake/Legitimate)**	**Avg. Word Count**	**Avg. Char Count**
FakeNewsNet	372 (171/201)	549	3087
ISOT	43,729 (21,416/22,313)	415	2514
FakeNewsKaggle	17,759 (7401/10,358)	524	4865
FakeNewsAMT	480 (240/240)	123	735
FakeNewsRandomPolitical	150 (75/75)	587	3611
FakeNewsCelebrity	500 (250/250)	432	2443
FakeNewsBuzfeedPolitical	101 (48/53)	936	5572
FakeNewsPolitFalse	274 (137/137)	579	3506
FakeNewsSatirical	360 (180/180)	543	3235

4.2. Preliminary Analysis

We conducted a preliminary analysis to gain insights into the most influential features of fake news detection. First, we conducted a feature correlation analysis to identify and eliminate highly correlated features across all datasets. We used a threshold value of 0.9 for the correlation coefficient, meaning that if two features correlated at 0.9 or higher, one was removed to reduce redundancy. As a result, the following readability features were removed due to high correlation: `RIX`, `characters_per_word`, `sentences`, `wordtypes`, `SMOGIndex`, `FleschReadingEase`, `syll_per_word`, `words`, `characters`, and `syllable`.

Next, we calculated the normalized mean value of each feature for both fake and legitimate news articles. By comparing these mean values, we could quantify the differences in how each feature appeared in fake versus legitimate news. To differentiate between fake and legitimate news, we calculate the difference between each feature for fake news and the corresponding values for legitimate news. A positive value from this calculation suggests a stronger association of that characteristic with fake content, while a negative value indicates a stronger association with legitimate content.

Figure 3 illustrates the features associated with a higher likelihood of fake news. The analysis reveals that fake news often exhibits high values in features such as `type_token_ratio`, which suggests that a diverse vocabulary may be employed to lend an appearance of sophistication or credibility. Additionally, the use of words expressing certainty (`liwc_certain`), present-tense language (`liwc_focuspresent`), and adverbs (`liwc_adverb`) are prevalent in fake news, reflecting a persuasive and often dramatic tone. Features like `liwc_function` and `liwc_pronoun` indicate that fake news tends to use more function words and pronouns, potentially to create detailed and relatable narratives. The frequent use of auxiliary verbs (`liwc_auxverb`), subjective language (`subjectivity`), and expressions of disgust (`emotion_disgust`) further highlights the manipulative nature of fake news, aiming to evoke strong emotional responses and sway reader opinions.

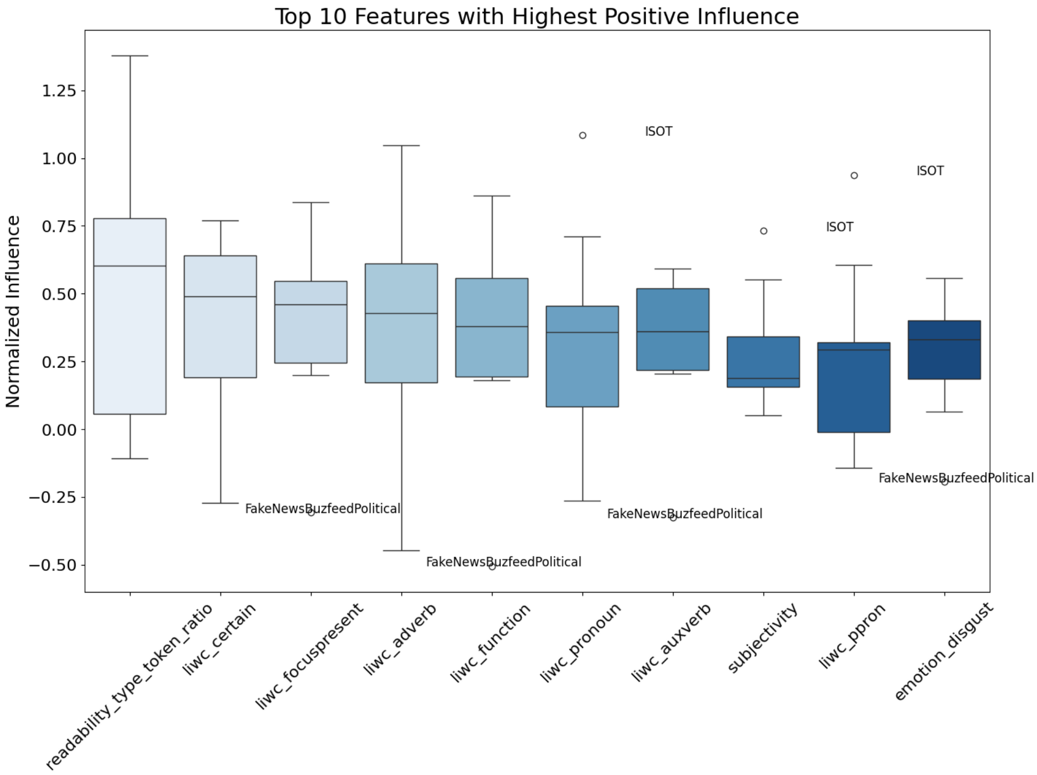

Figure 3. Top 10 features with higher values more commonly found in fake news.

Conversely, Figure 4 presents the features associated with a higher likelihood of legitimate news. Legitimate news articles are characterized by high values in features such as `liwc_hear`, reflecting a focus on information transmission through quoting sources and describing events. The use of complex words (`readability_complex_words` and `readability_complex_words_dc`), long words (`readability_long_words`), and nominalizations (`readability_nominalization`) indicates a sophisticated and formal language style typical of legitimate news. Furthermore, features like `readability_preposition`, `readability_article`, and `readability_sentences_per_paragraph` suggest that legitimate news employs clear, structured, and detailed writing. The use of subordinate clauses (`readability_subordination`) and work-related terms (`liwc_work`) also points to the thoroughness and factual nature of legitimate reporting, focusing on providing comprehensive and precise information.

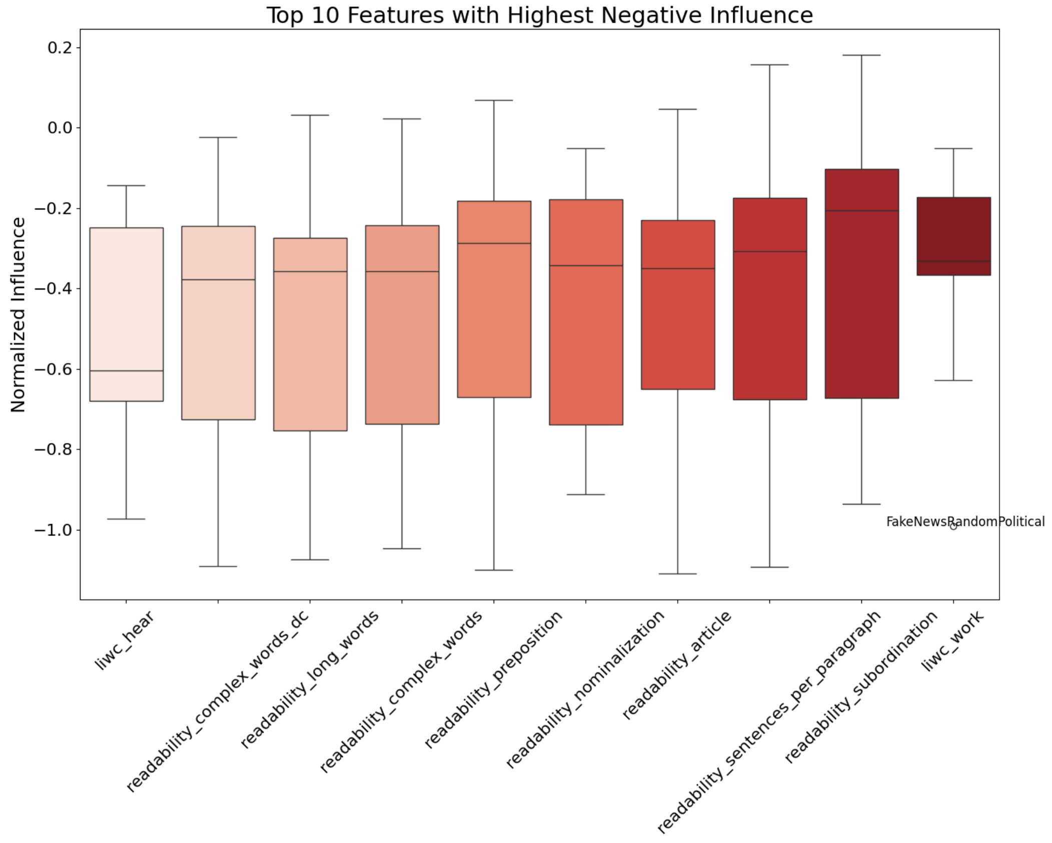

Figure 4. Top 10 features with higher values more commonly found in real news.

It is worth noting the significant variability in influence scores across datasets, as evidenced by the wide ranges in both figures. This variability underscores the complexity of fake news detection and the importance of considering multiple features and their interactions. The presence of outliers from specific datasets, such as ISOT and FakeNews-BuzfeedPolitical, highlights the importance of considering dataset-specific characteristics when targeting a specific domain or context.

Overall, the findings highlight that fake news is more likely to employ language that expresses certainty, focuses on the present, and uses a diverse vocabulary to appear sophisticated and credible. It also tends to include more adverbs, pronouns, and function words, contributing to a persuasive and often dramatic tone. In contrast, legitimate news is characterized by the use of complex but readable language. The presence of complex words, nominalizations, and structured writing with clear sentence construction underscores the thoroughness and factual nature of legitimate reporting. These differences underscore the manipulative nature of fake news, which seeks to evoke strong emotional responses, versus the detailed and precise information typical of legitimate news (**RQ1**).

4.3. Classification Performance

In this section, we describe the experiments conducted to evaluate the classification performance of the proposed model on the task of fake news detection. We utilize a variety of traditional classifiers, including logistic regression, support vector machines (SVM), decision trees, and ensemble methods like random forests, XGBoost, and CatBoost. We used default hyperparameter settings provided by their respective libraries for all classifiers.

Regarding performance metrics, we prioritized the most commonly used ones in related work: accuracy, precision, recall, and F1 score. Accuracy, specifically, measures the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances. It provides an overall assessment of the model's performance but can be less informative in cases of class imbalance. Precision measures the proportion of true positive predictions (TP) among all positive predictions, calculated as

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall (or sensitivity) measures the proportion of true positive predictions among all actual positives, calculated as

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F1 score is defined as the harmonic mean of precision and recall, which provides a single metric that balances both

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}$$

We employed cross-validation with $k = 5$ to ensure robust evaluation and applied standard scaling to normalize the features. Experiments have been implemented using Python Scikit-learn library [62] and run on a Dell Inc. XPS 13 9310 11th Gen i7 with 32GB RAM, except for experiments with transformers-based models, which have been run on a Kaggle notebook with GPU P100 (Available at https://www.kaggle.com/docs/notebooks, accessed on 9 July 2024).

In fake news detection, particularly when deployed under resource constraints, the model's efficiency can be as critical as its classification performance. This is especially relevant in scenarios where decisions must be made quickly and at scale. Therefore, in addition to evaluating the classification performance of different models, we considered training time a crucial factor in our evaluation [63]. As such, models that offer a good trade-off between accuracy and training time were prioritized for further analysis and discussion.

To further enhance the interpretability of our model, we conducted an analysis using SHAP (Shapley additive explanations) values [64]. SHAP is a well-established method whose theoretical foundation is based on cooperative game theory. It provides insights into the contribution of each feature to the model's predictions, offering a transparent view of how different features impact the outcomes. By applying SHAP, we can identify the most influential features in predicting outcomes and better understand the model's decision-making process. SHAP's advantage over traditional feature importance methods, such as feature permutation or information gain, lies in its fair and consistent feature attribution. Furthermore, SHAP is model-agnostic, making it more versatile than methods specific to certain model types.

For the first experiment, we trained the classifiers using the features extracted from the text (after removing highly correlated ones). The average results for all datasets grouped by the classifier are detailed in Table 2. In addition, Table 3 shows the results of each dataset of our most-relevant methods and compares them with state-of-the-art solutions.

Table 2. Performance metrics, training time, and accuracy of different algorithms. * BERT was run in a different environment than the other classifiers.

Algorithm	Train Time (s)	Accuracy	Precision	Recall	F1 Score
CatBoost	41.683	0.8146 ± 0.04	0.8184 ± 0.04	0.8129 ± 0.04	0.8122 ± 0.04
DecisionTree	0.846	0.7247 ± 0.04	0.7292 ± 0.04	0.7239 ± 0.03	0.7231 ± 0.03
LinearSVC	0.733	0.7539 ± 0.04	0.7575 ± 0.04	0.7541 ± 0.04	0.7533 ± 0.04
LogReg	0.172	0.7711 ± 0.04	0.7759 ± 0.04	0.7711 ± 0.04	0.7703 ± 0.04
RF	3.820	0.7895 ± 0.04	0.7949 ± 0.04	0.7890 ± 0.04	0.7875 ± 0.04
XGBoost	1.112	0.7933 ± 0.04	0.7991 ± 0.05	0.7931 ± 0.05	0.7919 ± 0.05
BERT	878.153 *	0.8136 ± 0.13	0.8512 ± 0.04	0.8287 ± 0.04	0.8247 ± 0.05

Table 3. Performance metrics by dataset using XGBoost, CatBoost, logistic regression, BERT, and state-of-the-art methods. * No approach using this dataset has been found in the literature.

Dataset	Method	Precision	Recall	F1 Score
FakeNewsNet	XGBoost	0.711648	0.709694	0.709650
	CatBoost	0.742586	0.739279	0.738630
	LogReg	0.660303	0.655892	0.654180
	BERT	0.732913	0.728505	0.723411
	SOTA [56]	0.671	0.738	0.703
ISOT	XGBoost	0.979265	0.979236	0.979234
	CatBoost	0.979627	0.979602	0.979600
	LogReg	0.964471	0.964463	0.964463
	BERT	0.998788	0.998788	0.998788
	SOTA [65]	0.9912	0.9914	0.9920
FakeNewsKaggle	XGBoost	0.880271	0.880455	0.880238
	CatBoost	0.881027	0.881187	0.880855
	LogReg	0.844767	0.842896	0.843373
	BERT	0.997077	0.997072	0.997071
	SOTA [66]	0.946	0.918	0.932
FakeNewsAMT	XGBoost	0.623906	0.622917	0.621729
	CatBoost	0.654660	0.652083	0.651102
	LogReg	0.640252	0.639583	0.639195
	BERT	0.714853	0.712500	0.710773
	SOTA [58]	0.75	0.74	0.74
FakeNewsRandomPolitical	XGBoost	0.810246	0.800000	0.798637
	CatBoost	0.805006	0.800000	0.799210
	LogReg	0.788832	0.786667	0.786087
	BERT	0.804778	0.780000	0.778705
	SOTA [48]	0.96	0.92	0.94
FakeNewsCelebrity	XGBoost	0.756078	0.754000	0.753462
	CatBoost	0.772602	0.768000	0.767128
	LogReg	0.685015	0.682000	0.680640
	BERT	0.806663	0.796000	0.793851
	SOTA [58]	0.73	0.73	0.73
FakeNewsBuzfeedPolitical	XGBoost	0.790064	0.762381	0.757566
	CatBoost	0.874191	0.851429	0.849208
	LogReg	0.829941	0.801905	0.798560
	BERT	0.607238	0.603333	0.560449
	SOTA [48]	1.00	0.83	0.90
FakeNewsPolitFalse	XGBoost	0.755926	0.748215	0.746870
	CatBoost	0.778025	0.773805	0.773279
	LogReg	0.699198	0.696970	0.696204
	BERT	0.808530	0.795758	0.790898
	SOTA *	-	-	-
FakeNewsSatirical	XGBoost	0.884115	0.880556	0.880124
	CatBoost	0.886940	0.886111	0.886028
	LogReg	0.870373	0.869444	0.869386
	BERT	0.913607	0.911111	0.910793
	SOTA [61]	0.88	0.82	0.87

CatBoost, though the slowest with a training time of 41.683 s, delivered the highest weighted F1 score of 0.8122, demonstrating its strong classification performance. In contrast, DecisionTree, with its rapid fit time of 0.846 s, produced the lowest F1 score of 0.7231. XGBoost, with a training time of 1.112 s, achieved an F1 score of 0.7919, striking a commendable balance between performance and computational efficiency. Although CatBoost yielded the highest performance, its substantially higher training time limits its practical application in scenarios requiring rapid processing. In this regard, logistic regression showed the lowest training time. XGBoost, on the other hand, offers a solid compromise between classification performance and training time. The final choice between these classifiers would depend on the specific requirements of the application, such as the need for real-time processing or the availability of computational resources.

As we can observe, our approach delivers highly competitive performance while drastically reducing training time compared to transformer-based models. Although BERT slightly edges out in terms of F1 score, it comes with a significant computational cost, requiring around 878 s to train, while CatBoost and XGBoost only take 4.75% and 0.13% of that time, respectively. This demonstrates that while transformer-based models offer strong results, our approach yields competitive performance at a fraction of the computational time (**RQ2**).

Another interesting aspect of the proposed solution is that it demonstrates strong performance across various datasets, mainly when using ensemble algorithms. It produced auspicious results in the ISOT, FakeNewsKaggle, and FakeNewsSatirical datasets. Generally, precision and recall are closely aligned, and most datasets show minimal differences between the two metrics. The lowest scores were observed in the FakeNewsAMT dataset, likely due to the generally shorter news articles in this dataset, as indicated by the statistics in Table 1. Shorter articles may provide less information, impacting the model's performance.

Although state-of-the-art methods generally outperform it, our approach achieves competitive performance while preserving significant advantages regarding computational efficiency and explainability. Furthermore, our method demonstrates robust performance across various datasets, validating its generalizability and practical applicability.

To further understand the impact of feature reduction on classification performance, we conducted a series of experiments by incrementally reducing the number of features used by traditional classifiers. The selection process for the feature set was based on the preliminary analysis described in Section 4.2, which identified the features with the most significant differences between fake and legitimate news. Specifically, we selected the 25 features whose average values were significantly higher in fake news and the 25 features whose average values were significantly higher in legitimate news. Starting with this set of 50 features, we incrementally reduced the number of features used in our models, evaluating performance at each step.

Figure 5 illustrates the relationship between the number of features and the model's performance, measured using the weighted F1 score. The results show that the average performance across datasets stabilizes between 20 and 30 features, with an average F1 score of around 0.75. However, the trend varies between datasets. For instance, performance in ISOT and FakeNewsKaggle datasets continues to increase and only tends to stabilize after 40 features. The large size and diversity of these datasets could explain this matter. In contrast, performance in FakeNewsBuzfeedPolitical and FakeNewsRandomPolitical peaks before reaching 20 features and tends to drop after 30 features.

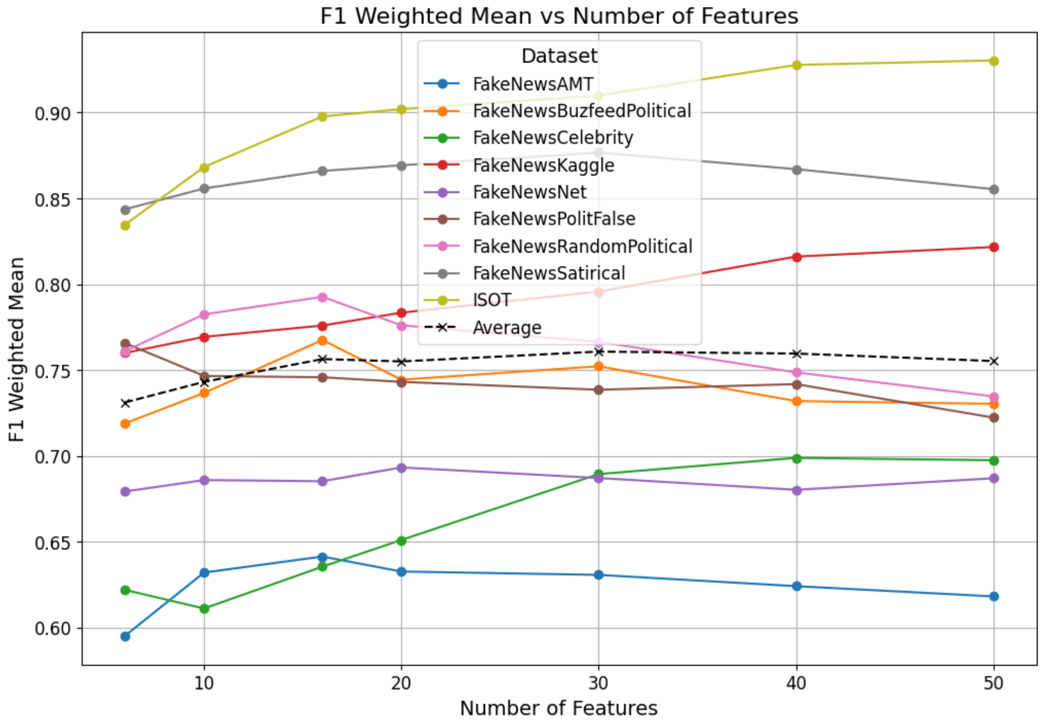

Figure 5. F1 score variation with feature count across different datasets.

In terms of computational efficiency, reducing the number of features has a substantial impact on training time. As shown in Table 4, all algorithms tested experienced substantial decreases in training time when using a reduced feature set of 30 features. On average, the algorithms saw their training times reduced by more than 70%. Also, it is worth noting that reducing the number of features not only enhances efficiency but also contributes to making the model more interpretable. In general, these findings suggest that reducing the number of features to around 30 can significantly improve the models' efficiency and interpretability without seriously compromising performance (**RQ3**). Although some datasets, like ISOT and FakeNewsKaggle, benefit from a larger feature set, most datasets achieve near-optimal performance with fewer features. Limiting the feature set reduces model complexity, leading to faster training times and simplifying the model's structure. This makes understanding and interpreting the relationships between the features and the predictions easier.

Table 4. Training time of different algorithms with reduction percentage. The new training time is measured when training models with 30 features.

Algorithm	Original Training Time (s)	New Training Time (s)	Reduction (%)
CatBoost	41.683	12.912	69.03%
DecisionTree	0.846	0.229	72.94%
LinearSVC	0.733	0.078	89.36%
LogisticRegression	0.172	0.026	84.77%
RandomForest	3.820	1.968	48.52%
XGBoost	1.112	0.308	72.29%

Finally, we conducted an analysis using SHAP to further enhance the interpretability of our model. In this analysis, we focused on the ten features with the highest positive influence and those with the highest negative influence on the model's predictions. Given its consistent balance between accuracy and speed, we used XGBoost as the method for this study. In addition, we have aggregated the SHAP values across all datasets to achieve a broad understanding.

Figure 6 shows the results of the experiment. The features are ranked by their average impact on the model's output magnitude. The most influential feature appears to be `liwc_hear`, followed closely by `liwc_pronoun`. Several linguistic-related features like `readability_article`, `readability_complex_words_dc`, and `readability_long_words` also show significant influence. Other characteristics such as adverbs, focus on the present, and certainty markers have moderate impacts. Finally, some affective features like `subjectivity` and `emotion_disgust` appear lower on the list, suggesting they have a smaller but still noticeable effect on the model's output.

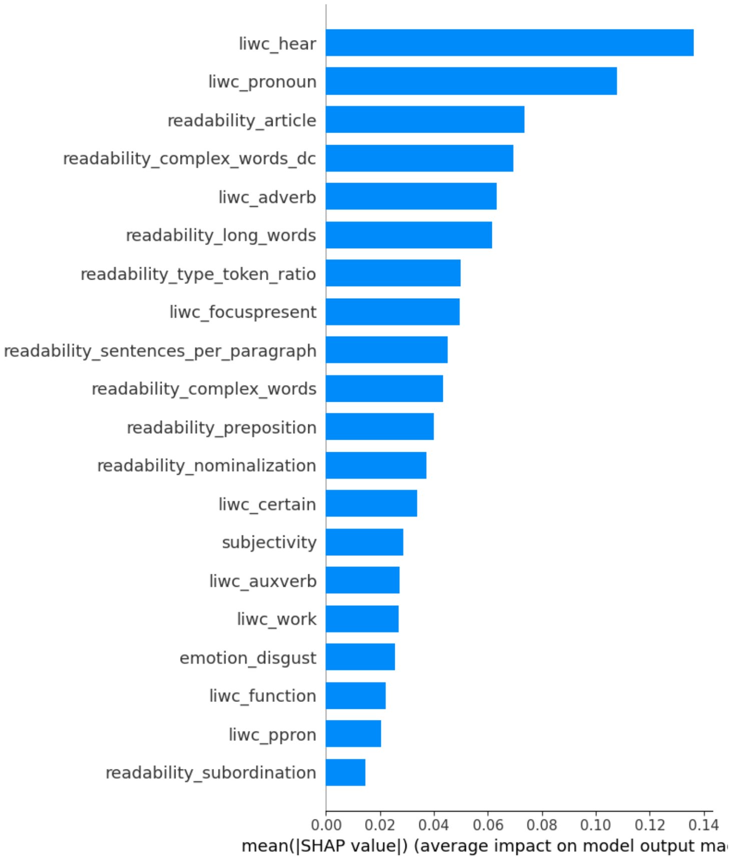

Figure 6. Ranking of features according to their impact on model output.

5. Discussion

In this study, we have delved into the critical issue of fake news detection by developing and evaluating a comprehensive feature framework. We have made three primary contributions: First, we introduced a novel set of features that capture subtle content characteristics indicative of fake news. Second, our evaluation using multiple traditional classifiers demonstrated high predictive performance of these features. Finally, we enhanced the efficiency and interpretability of our model by analyzing the effects of feature set reduction, optimizing both performance and clarity in the detection process. The conducted explainability analysis provided additional insights into the decision-making process of our detection system. Overall, our findings advance the field of fake news detection by providing practical, interpretable solutions that improve accuracy and offer a foundation for future research.

Our analysis has illuminated significant distinctions between fake and legitimate news based on feature influence. Fake news often employs language that conveys certainty, uses diverse vocabulary, and includes frequent adverbs, pronouns, and function words to create a persuasive and dramatic tone. In contrast, legitimate news is characterized by complex and structured language, including complex words, nominalizations, and clear sentence construction, reflecting thoroughness and factual reporting (**RQ1**).

The machine learning model exploiting the proposed framework showed significant performance. The evaluation of classifiers revealed that our approach offered a strong balance between accuracy and efficiency, showcasing its effectiveness. Additionally, its results were competitive with transformer-based models like BERT (**RQ2**). The investigation into feature reduction demonstrated that including up to 30 features optimizes the relation between performance and efficiency (**RQ3**). The SHAP analysis further enhanced our understanding of feature contributions, providing valuable insights into model interpretability. This understanding underscores the manipulative nature of fake news and the sophisticated characteristics of legitimate news.

The findings of this work are expected to provide valuable insights into the role of interpretable features in fake news detection and offer practical solutions for developing more accessible and transparent detection systems. Focusing on content characteristics provides a viable alternative to resource-intensive transformer-based models, which are often costly in terms of computational power and energy consumption. Our feature-based framework can be more easily integrated into existing fake news detection systems, particularly in environments with limited computational resources.

In terms of practical implementation, our approach could be embedded into content moderation tools used by social media platforms, providing interpretable decisions on whether an article is likely fake or legitimate. Additionally, it could be adapted into browser extensions or AI-powered fact-checking tools that assist users in assessing the credibility of online content. These tools would not only serve journalists and fact-checkers but could also enhance public media literacy by providing real-time feedback to users about potentially misleading information.

However, several challenges may arise when practitioners attempt to apply our framework in real-world settings. First, the performance of our feature-based model is closely tied to the quality and completeness of the data available. In environments where content characteristics differ significantly from those of the datasets we studied (e.g., non-English content, multimedia posts, or nuanced satire), additional feature engineering may be necessary to ensure robustness. Another challenge is the potential evolution of fake news tactics. As fake news creators become more sophisticated, new content features may emerge, requiring continuous updates to the feature sets used by the detection models.

Furthermore, this study has several limitations that should be acknowledged. Firstly, the reliance on traditional classifiers and the feature sets employed may only partially capture the nuances of more sophisticated models, potentially limiting the scope of our findings. Although we compared our methods to transformer-based models, the focus on BERT might overlook advancements in deep learning techniques. Additionally, the dependence on English-language datasets also restricts the generalizability of our findings to other languages and cultural contexts.

Several avenues for future research could further advance the field and help address current limitations. Expanding the study to include multiple languages and cultural contexts could lead to the development of more globally applicable detection methods. This could involve creating multilingual datasets and adapting our feature framework to capture language-specific nuances. Additionally, expanding the feature set to include additional characteristics could further improve detection performance, making the models more robust and adaptable. Also, integrating these methods into real-world environments is crucial for evaluating their practicality and effectiveness in diverse contexts. Such integration will help develop more transparent, efficient, and practical fake news detection systems. Ultimately, this will support the preservation of information integrity and enhance the reliability of news sources in an increasingly complex digital landscape. Finally, conducting in-depth studies on the ethical implications of automated fake news detection, including potential biases in training data and model decisions, is essential. Developing methods to mitigate these biases and ensure fair application across diverse populations will be critical for the responsible deployment of these technologies.

Our work provides a foundation for these future endeavors, offering insights into the role of interpretable features in fake news detection and presenting practical solutions for developing more accessible and transparent detection systems.

Author Contributions: Conceptualization, S.M. and C.A.I.; methodology, S.M. and C.A.I.; software, S.M.; writing—original draft preparation, S.M.; writing—review and editing, S.M. and C.A.I.; visualization, S.M.; project administration, C.A.I.; funding acquisition, C.A.I. All authors have read and agreed to the published version of the manuscript.

Funding: This project has been funded by the project UNICO I+D Cloud-AMOR, financed by the Ministry of Economic Affairs and Digital Transformation and the European Union through Next Generation EU.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available in a publicly accessible repository at https://github.com/gsi-upm/explainable-fn-detection accessed on 15 June 2024.

Acknowledgments: Authors would like to thank the researchers who collected and published the used datasets. Author Sergio Muñoz would like to thank the TAILOR project (funded by the EU Horizon 2020 research and innovation program under GA No. 952215) for the connectivity fund which provided knowledge that contributed to the completion of this article. Figure 1 is based on a template by PresentationGO (Available at www.presentationgo.com, accessed on 23 July 2024) and icons are designed by Freepik (Available at www.freepik.es, accessed on 23 July 2024).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

LIWC	Linguistic inquiry and word count
RQ	Research question
SHAP	Shapley additive explanations
VADER	Valence-Aware Dictionary and sEntiment Reasoner)

References

1. Parikh, S.B.; Patil, V.; Atrey, P.K. On the origin, proliferation and tone of fake news. In Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 28–30 March 2019; pp. 135–140.
2. Sumpter, M.; Ciampaglia, G.L. Preserving the Integrity and Credibility of the Online Information Ecosystem. *IEEE Data Eng. Bull.* **2021**, *44*, 4–11.
3. Molina, M.D.; Sundar, S.S.; Le, T.; Lee, D. "Fake news" is not simply false information: A concept explication and taxonomy of online content. *Am. Behav. Sci.* **2021**, *65*, 180–212. [CrossRef]
4. Baptista, J.P.; Gradim, A. A working definition of fake news. *Encyclopedia* **2022**, *2*, 632–645. [CrossRef]
5. Tsfati, Y.; Boomgaarden, H.G.; Strömbäck, J.; Vliegenthart, R.; Damstra, A.; Lindgren, E. Causes and consequences of mainstream media dissemination of fake news: Literature review and synthesis. *Ann. Int. Commun. Assoc.* **2020**, *44*, 157–173. [CrossRef]
6. Mwangi, M. Technology and Fake News: Shaping Social, Political, and Economic Perspectives. *Biomed. Sci. Clin. Res.* **2023**, *2*, 221–236. [CrossRef]
7. Karppi, T.; Crawford, K. Social media, financial algorithms and the hack crash. *Theory Cult. Soc.* **2016**, *33*, 73–92. [CrossRef]
8. Rocha, Y.M.; De Moura, G.A.; Desidério, G.A.; De Oliveira, C.H.; Lourenço, F.D.; de Figueiredo Nicolete, L.D. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *J. Public Health* **2021**, *31*, 1007–1016. [CrossRef] [PubMed]
9. Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **2017**, *31*, 211–236. [CrossRef]
10. Calvillo, D.P.; Rutchick, A.M.; Garcia, R.J. Individual differences in belief in fake news about election fraud after the 2020 US election. *Behav. Sci.* **2021**, *11*, 175. [CrossRef] [PubMed]
11. Barrera, O.; Guriev, S.; Henry, E.; Zhuravskaya, E. Facts, alternative facts, and fact checking in times of post-truth politics. *J. Public Econ.* **2020**, *182*, 104123. [CrossRef]
12. Mutahi, P. Fake news and the 2017 Kenyan elections. *Commun. S. Afr. J. Commun. Theory Res.* **2020**, *46*, 31–49. [CrossRef]

13. Airlangga, G. Comparative Analysis of Machine Learning Algorithms for Detecting Fake News: Efficacy and Accuracy in the Modern Information Ecosystem. *J. Comput. Netw. Archit. High Perform. Comput.* **2024**, *6*, 354–363. [CrossRef]
14. Al-Asadi, M.A.; Tasdemir, S. Using artificial intelligence against the phenomenon of fake news: A systematic literature review. *Combat. Fake News Comput. Intell. Tech.* **2022**, *1001*, 39–54.
15. Kapusta, J.; Benko, L.; Munk, M. Fake news identification based on sentiment and frequency analysis. In *Proceedings of the Innovation in Information Systems and Technologies to Support Learning Research*; Proceedings of EMENA-ISTL 2019; Springer: Cham, Switzerland, 2020; pp. 400–409.
16. Choudhary, A.; Arora, A. Linguistic feature based learning model for fake news detection and classification. *Expert Syst. Appl.* **2021**, *169*, 114171. [CrossRef]
17. Verma, P.K.; Agrawal, P.; Amorim, I.; Prodan, R. WELFake: Word embedding over linguistic features for fake news detection. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 881–893. [CrossRef]
18. Mridha, M.F.; Keya, A.J.; Hamid, M.A.; Monowar, M.M.; Rahman, M.S. A comprehensive review on fake news detection with deep learning. *IEEE Access* **2021**, *9*, 156151–156170. [CrossRef]
19. Fields, J.; Chovanec, K.; Madiraju, P. A survey of text classification with transformers: How wide? How large? How long? How accurate? How expensive? How safe? *IEEE Access* **2024**, *12*, 6518–6531. [CrossRef]
20. Azizah, S.F.N.; Cahyono, H.D.; Sihwi, S.W.; Widiarto, W. Performance Analysis of Transformer Based Models (BERT, ALBERT, and RoBERTa) in Fake News Detection. In Proceedings of the 2023 6th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 10 November 2023; pp. 425–430.
21. Naseer, M.; Windiatmaja, J.H.; Asvial, M.; Sari, R.F. RoBERTaEns: Deep Bidirectional Encoder Ensemble Model for Fact Verification. *Big Data Cogn. Comput.* **2022**, *6*, 33. [CrossRef]
22. Patwardhan, N.; Marrone, S.; Sansone, C. Transformers in the real world: A survey on NLP applications. *Information* **2023**, *14*, 242. [CrossRef]
23. Bakir, V.; McStay, A. Fake news and the economy of emotions: Problems, causes, solutions. *Digit. J.* **2018**, *6*, 154–175. [CrossRef]
24. Hu, L.; Wei, S.; Zhao, Z.; Wu, B. Deep learning for fake news detection: A comprehensive survey. *AI Open* **2022**, *3*, 133–155. [CrossRef]
25. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
26. Drif, A.; Hamida, Z.F.; Giordano, S. Fake news detection method based on text-features. *Fr. Int. Acad. Res. Ind. Assoc. (IARIA)* **2019**, 27–32.
27. Przybyla, P. Capturing the style of fake news. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 490–497.
28. Cao, J.; Qi, P.; Sheng, Q.; Yang, T.; Guo, J.; Li, J. Exploring the role of visual content in fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 141–161.
29. Shu, K.; Wang, S.; Liu, H. Beyond news contents: The role of social context for fake news detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, New York, NY, USA, 11–15 February 2019; pp. 312–320.
30. Reis, J.C.; Correia, A.; Murai, F.; Veloso, A.; Benevenuto, F. Supervised learning for fake news detection. *IEEE Intell. Syst.* **2019**, *34*, 76–81. [CrossRef]
31. Hauschild, J.; Eskridge, K. Word embedding and classification methods and their effects on fake news detection. *Mach. Learn. Appl.* **2024**, *17*, 100566. [CrossRef]
32. Muñoz, S.; Iglesias, C.A. A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. *Inf. Process. Manag.* **2022**, *59*, 103011. [CrossRef]
33. Muñoz, S.; Iglesias, C.Á. Detection of the Severity Level of Depression Signs in Text Combining a Feature-Based Framework with Distributional Representations. *Appl. Sci.* **2023**, *13*, 11695. [CrossRef]
34. Ge, X.; Hao, S.; Li, Y.; Wei, B.; Zhang, M. Hierarchical co-attention selection network for interpretable fake news detection. *Big Data Cogn. Comput.* **2022**, *6*, 93. [CrossRef]
35. Sastrawan, I.K.; Bayupati, I.P.A.; Arsa, D.M.S. Detection of fake news using deep learning CNN–RNN based methods. *ICT Express* **2022**, *8*, 396–408. [CrossRef]
36. Goonathilake, M.P.; Kumara, P.V. CNN, RNN-LSTM based hybrid approach to detect state-of-the-art stance-based fake news on social media. In Proceedings of the 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 4–7 November 2020; pp. 23–28.
37. Comito, C.; Caroprese, L.; Zumpano, E. Multimodal fake news detection on social media: A survey of deep learning techniques. *Soc. Netw. Anal. Min.* **2023**, *13*, 101. [CrossRef]
38. Ma, Z.; Luo, M.; Guo, H.; Zeng, Z.; Hao, Y.; Zhao, X. Event-Radar: Event-driven Multi-View Learning for Multimodal Fake News Detection. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 5809–5821.
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.

40. Kaliyar, R.K.; Goswami, A.; Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimed. Tools Appl.* **2021**, *80*, 11765–11788. [CrossRef] [PubMed]
41. Almaliki, M.; Almars, A.M.; Gad, I.; Atlam, E.S. Abmm: Arabic bert-mini model for hate-speech detection on social media. *Electronics* **2023**, *12*, 1048. [CrossRef]
42. Farhangian, F.; Cruz, R.M.; Cavalcanti, G.D. Fake news detection: Taxonomy and comparative study. *Inf. Fusion* **2024**, *103*, 102140. [CrossRef]
43. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. [CrossRef]
44. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
45. Reis, J.C.; Correia, A.; Murai, F.; Veloso, A.; Benevenuto, F. Explainable machine learning for fake news detection. In Proceedings of the 10th ACM Conference on Web Science, New York, NY, USA, 30 June 2019–3 July 2019; pp. 17–26.
46. Liu, H.; Wang, W.; Li, H. Interpretable Multimodal Misinformation Detection with Logic Reasoning. *arXiv* **2023**, arXiv:2305.05964.
47. Han, L.; Zhang, X.; Zhou, Z.; Liu, Y. A Multifaceted Reasoning Network for Explainable Fake News Detection. *Inf. Process. Manag.* **2024**, *61*, 103822. [CrossRef]
48. Garg, S.; Sharma, D.K. Linguistic features based framework for automatic fake news detection. *Comput. Ind. Eng.* **2022**, *172*, 108432. [CrossRef]
49. Kondamudi, M.R.; Sahoo, S.R.; Chouhan, L.; Yadav, N. A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. *J. King Saud-Univ.-Comput. Inf. Sci.* **2023**, *35*, 101571. [CrossRef]
50. Hu, B.; Mao, Z.; Zhang, Y. An Overview of Fake News Detection: From A New Perspective. *Fundam. Res.* **2024**, in press.
51. Shrestha, A.; Spezzano, F. Textual characteristics of news title and body to detect fake news: A reproducibility study. In *Proceedings of the Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, 28 March–1 April 2021*; Proceedings, Part II 43; Springer: Berlin/Heidelberg, Germany, 2021; pp. 120–133.
52. Carrasco-Farré, C. The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanit. Soc. Sci. Commun.* **2022**, *9*, 162. [CrossRef]
53. Pennebaker, J.W.; Francis, M.E.; Booth, R.J. Linguistic inquiry and word count: LIWC 2001. *Mahway Lawrence Erlbaum Assoc.* **2001**, *71*, 2001.
54. Araque, O.; Gatti, L.; Kalimeri, K. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowl.-Based Syst.* **2020**, *191*, 105184. [CrossRef]
55. Ahmed, H.; Traore, I.; Saad, S. Detecting opinion spams and fake news using text classification. *Secur. Priv.* **2018**, *1*, e9. [CrossRef]
56. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv* **2018**, arXiv:1809.01286. [CrossRef] [PubMed]
57. Lifferth, W. Fake News, 2018. Available online: https://www.kaggle.com/competitions/fake-news/overview (accessed on 12 June 2024).
58. Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; Mihalcea, R. Automatic Detection of Fake News. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 3391–3401.
59. Horne, B.; Adali, S. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; Volume 11, pp. 759–766.
60. Asubiaro, T.V.; Rubin, V.L. Comparing features of fabricated and legitimate political news in digital environments (2016–2017). *Proc. Assoc. Inf. Sci. Technol.* **2018**, *55*, 747–750. [CrossRef]
61. Rubin, V.L.; Conroy, N.; Chen, Y.; Cornwell, S. Fake news or truth? Using satirical cues to detect potentially misleading news. In Proceedings of the Second Workshop on Computational Approaches to Deception Detection, San Diego, CA, USA, 17 June 2016; pp. 7–17.
62. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
63. Khan, J.Y.; Khondaker, M.T.I.; Afroz, S.; Uddin, G.; Iqbal, A. A benchmark study of machine learning models for online fake news detection. *Mach. Learn. Appl.* **2021**, *4*, 100032. [CrossRef]
64. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
65. Nadeem, M.I.; Mohsan, S.A.H.; Ahmed, K.; Li, D.; Zheng, Z.; Shafiq, M.; Karim, F.K.; Mostafa, S.M. HyproBert: A fake news detection model based on deep hypercontext. *Symmetry* **2023**, *15*, 296. [CrossRef]
66. Parmar, S.; Rahul. Fake news detection via graph-based Markov chains. *Int. J. Inf. Technol.* **2024**, *16*, 1333–1345. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.