electronics

MDPI

Article

An Explainable Fake News Analysis Method with Stance Information

Lu Yuan [1,2], Hao Shen [2], Lei Shi [2,*], Nanchang Cheng [2,*] and Hangshun Jiang [2]

[1] School of Data Science and Media Intelligence, Communication University of China, Beijing 100024, China; yuanlucuc@cuc.edu.cn
[2] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China; shenhao@cuc.edu.cn (H.S.); 311909011312@home.hpu.edu.cn (H.J.)
* Correspondence: leiky_shi@cuc.edu.cn (L.S.); chengnanchang@cuc.edu.cn (N.C.)

Abstract: The high level of technological development has enabled fake news to spread faster than real news in cyberspace, leading to significant impacts on the balance and sustainability of current and future social systems. At present, collecting fake news data and using artificial intelligence to detect fake news have an important impact on building a more sustainable and resilient society. Existing methods for detecting fake news have two main limitations: they focus only on the classification of news authenticity, neglecting the semantics between stance information and news authenticity. No cognitive-related information is involved, and there are not enough data on stance classification and news true-false classification for the study. Therefore, we propose a fake news analysis method based on stance information for explainable fake news detection. To make better use of news data, we construct a fake news dataset built on cognitive information. The dataset primarily consists of stance labels, along with true-false labels. We also introduce stance information to further improve news falsity analysis. To better explain the relationship between fake news and stance, we use propensity score matching for causal inference to calculate the correlation between stance information and true-false classification. The experiment result shows that the propensity score matching for causal inference yielded a negative correlation between stance consistency and fake news classification.

Keywords: stance information; fake news analysis; explainable AI system; PSM

Citation: Yuan, L.; Shen, H.; Shi, L.; Cheng, N.; Jiang, H. An Explainable Fake News Analysis Method with Stance Information. *Electronics* **2023**, *12*, 3367. https://doi.org/10.3390/electronics12153367

Academic Editor: Arkaitz Zubiaga

Received: 4 July 2023
Revised: 22 July 2023
Accepted: 4 August 2023
Published: 7 August 2023

1. Introduction

A large amount of fake news spreads on the Internet, giving rise to an information epidemic that significantly impacts the balance and sustainability of the current and future social systems. According to a survey of current fake news detection methods, deep learning models are commonly used for semantic feature extraction [1–3]. In addition, fake news detection can also be accomplished through knowledge augmentation combined with the consideration of user comments [4–6]. Furthermore, fake news has complex underlying reasons, underscoring the importance of its interpretability in the detection process. Not only does this help gain a high degree of trust from the audience through explanations, but it also effectively improves detection performance by continuously optimizing explanations [7]. Although the research field of explainable fake news detection has achieved some results, it is still far from meeting expectations.

Among the intricate elements of news, the stance information of the news subject has become a critical factor in judging the credibility of the news. When analyzing fake news through stance detection, the combination of stance detection with knowledge and deep learning method modeling enables effective analysis [8,9]. Numerous studies have demonstrated the effectiveness of incorporating stance information into the detection and analysis of fake news [10,11]. However, the current fake news detection methods with stance information are still limited in performance and face the following challenges:

(1) Lack of task-related datasets. Existing methods strive to make full use of public datasets but lack stance information data that exclude cognitive information related to partisanship and bias. In addition, fake news also suffers from serious imbalance problems in the real world, and it is necessary to further explore how to effectively utilize existing data with data as the center to improve the accuracy of fake news detection.
(2) Lack of stance information application. The current fake news detection work only distinguishes the authenticity of fake news detection from the level of semantic information. Although the fake news detection model can provide the detection results of news authenticity, users often do not know whether such detection results are reliable. Previous studies have solely focused on technological innovations in the aspect of authenticity identification of fake news without considering stance information for explainable fake news analysis.

In this paper, we aim to tackle the above issues by introducing a fake news analysis method: Stance Classification Fake News Analysis (SC-FNA). The advantages of SC-FNA are three-fold: (1) Utilizing stance information: we combine cognitive information and surveys to construct a fake news dataset, which has both stance labels and true-false labels. (2) Addressing imbalanced data: to overcome the problem of imbalanced data classification, we propose an integrated data augmentation method to form an extended dataset for research. (3) Explaining the relationship between fake news and stance information: we use propensity score matching (PSM) for causal inference to calculate this relationship.

The main contributions of our paper are summarized as follows:

(1) Fake News Dataset with Stance Classification: We build a fake news dataset with both stance classification and fake or real classification. Through manual annotation, we use cognitive questionnaire surveys and mathematical modeling to integrate annotations for controlling the external factors of artificial bias.
(2) Integrated Data Augmentation Algorithm: We propose an integrated data augmentation algorithm. We use the existing data augmentation algorithms to combine and compare, and then explore the data augmentation algorithm combination that can best improve the accuracy performance.
(3) Explainable Fake News Analysis Method with Stance Information: We propose a fake news analysis method with stance information to form an explainable artificial intelligence system. We use the propensity score matching method to perform causal inference on the fake news and calculate the correlation between fake news classification and stance consistency.

The detailed chapters are arranged as follows: Section 2 is the related work. Section 3 is the FORSD dataset with both stance classification and true or false classification. Section 4 is the problem statement. Section 5 is the fake news analysis method with stance information. Section 6 is the experimental results and analysis. Section 7 is the conclusion.

2. Background and Related Work

2.1. Fake News Detection and Stance Detection

With the development of intelligent media and under the influence of the information epidemic, the spread of fake news has a significant negative impact on both society and individuals. From the perspective of specific research tasks, fake news detection can be divided into different subtasks: stance detection, topic detection, and fake news analysis [12]. As a crucial subtask, stance detection plays a vital role in extracting authenticity clues for identifying fake news [13]. Specifically, natural language processing technology is utilized for stance detection to assess the consistency of stance expressions in news text [14].

In 2017, the FNC-1 Fake News Challenge was launched, making stance detection a key initial step in fake news detection [15]. Stance detection essentially refers to the attitude expressed in text data toward a specific target, such as an event, person, or policy [16]. Depending on the target type, stance detection can be further categorized into single-target stance detection, multi-target position detection, and cross-target position detection [17].

In addition, based on text granularity, it can be divided into sentence-level and chapter-level stance detection [18].

In specific research endeavors, the methods for stance detection in fake news detection are continuously improving with the development of natural language processing technology. Noteworthy methodologies include stance detection based on traditional machine learning techniques, such as decision trees and Naive Bayes, to select appropriate models for feature representation [19]. Alternatively, considering the relationship between emotions and selected targets, stance detection can be jointly accomplished with the semantic representation of emotional features [20]. Moreover, stance detection based on deep learning effectively captures grammar and syntactic information for accurate stance detection [21]. Finally, utilizing pre-trained models like BERT, GPT, and other pre-trained language models for semantic representation enables the accomplishment of the final stance detection task through fine-tuning techniques [13]. At present, stance detection is performed through small-dsample and multi-task learning [22,23], and there are also promising studies on stance detection using large language models [24].

2.2. Explainable Fake News Detection

Explainable AI systems for natural language processing can be divided into interpretable model structures and interpretable model behaviors. For explainable fake news detection, it can also be roughly divided into two categories. The interpretable model structure analysis is to analyze and understand the internal structure of the model through interpretable technology and to understand the working principle and working mechanism of the model. Song [25] proposed XFlag, an explainable AI (XAI) framework, which used LSTM for a fake news detection model and used the Layered Relevance Propagation (LRP) algorithm to explain the model. Wu [26] used a knowledge graph-enhanced representation learning framework for embeddings to detect fake news. Yu [27] used a multidisciplinary language synthesis method to train features that humans can understand and then used these features to train a deep learning classifier with a bidirectional recurrent neural network (BRNN) structure to make the classifier interpretable in news data, leading to stronger detection results.

The interpretable model behavior analysis involves performing interpretable analysis on the results predicted by the model and providing the basis for the predicted results. Yi et al. [28] utilized the Graph-aware Common Attention Network (GCAN) to judge the authenticity of source tweets in social media and provide explanations for the results. Hai et al. [29] proposed an automated interpretable decision system, QA-AXDS, based on quantitative argumentation, which can provide users with explanations about the results. Ni et al. [30] used the Multi-View Attention Network (MVAN) to detect fake news in social networks and provide an explanation for the results. In order to reduce the risk brought by the spread of fake news, in addition to the interpretability analysis of the model structure through machine learning, it is also necessary to carry out an explainable analysis of the model behavior of fake news. At present, whether it is from the analysis of model structure or from the analysis of model behavior, explainable artificial intelligence methods have not formed a complete solution.

At present, most of the existing methods have limitations. The existing research has only focused on one of the stance classification or authenticity classification of news while ignoring the relationship between them, resulting in the insufficient interpretability of fake news analysis.

3. FORSD: A Dataset for Fake News Stance Information

To enable our study of stance classification and to facilitate further research in explainable fake news analysis, we created a new available dataset that includes annotations for stance judgments.

3.1. Extracting Data for the Dataset

The primary dataset used for the labeling work is the "Fake_or_real_news" dataset published on Kaggle, which has been commonly used in previous studies for fake news detection applications.

This dataset consists of news data from two mainstream media sources, containing both true and fake news articles. Each data entry content includes the news ID, title (title), text (text), and true/false labels (label). The news topics cover political news, as well as social and technology news. The dataset contains a total of 6335 pieces of data, with an equal ratio of true and false news articles (1:1). Before commencing the dataset labeling work, we randomly sampled 2100 pieces of data by random sampling from the overall dataset. At the same time, to maintain fairness in the subsequent fake news analysis research, the ratio of true to false data of 2100 pieces of data was 1:1. As for the subject matter of news material, all political news material in the data was selected.

3.2. Labels for the Dataset

In terms of the selection of annotators, the study initially recruited 76 annotators. Before commencing data labeling, we conducted relevant demographic statistics. Specifically, in demographic statistics, in order to avoid gender bias when recruiting labelers, we ensured a balanced ratio of male to female annotators, maintaining a 1:1 proportion. In addition, for education statistics, we maintained an equal ratio of undergraduate to graduate students, also at 1:1. Finally, in view of the dataset's focus on political news and the sensitivity of the relevant political ideology, we conducted a partisan bias cognitive test of the annotators, In addition, a questionnaire was distributed to investigate any political bias. After the first round of questionnaires and modeling, we took measures to rule out obvious partisanship. Subsequently, we carefully screened the annotators and ultimately selected 60 volunteers who were best suited for the labeling task. The final selection maintained an equal 1:1 ratio of males to females and an equal 1:1 ratio of undergraduate to graduate education.

During the specific annotation work process, we adhered to the same stance definition as utilized in the FNC-1 dataset. Consequently, for each data instance, the stance judgment was categorized into four distinct classes, as presented in Table 1.

Table 1. The definition of stance.

Stance	Definition
Agree	The stance of the body and the title is the same
Disagree	Inconsistent stance between body and title
Unrelated	The content of the body and title is irrelevant
Discusses	The main body and the title express the same theme, but there is no clear stance

The dataset annotation work in this paper draws inspiration from Luo's research [31]. The specific process was as follows: 8 labels were collected for each piece of data, and a total of 16,800 labels were obtained. Similarly, during the labeling process, we found that the labels of some labelers were more reliable, so we chose to use the Bayesian model to aggregate the annotations of each data, and we assigned the label with the highest probability to each data on the basis of the Bayesian model. At the same time, in order to verify the consistency of data annotation by different annotators, the Kappa coefficient was used to measure consistency.

Manual labeling was conducted, but it was constrained by the limitations of the dataset itself and the experimental conditions, resulting in a more serious classification balance problem. The Kappa coefficient was introduced to perform the consistency check. During the experiment, the dataset was corrected according to the existing evaluation results, and

any incorrect labels in the existing manual labeling indicators were manually corrected. Based on the confusion matrix, the *kappa* coefficient was calculated as follows.

$$kappa = \frac{u_0 - u_a}{1 - u_a} \tag{1}$$

where u_0 is equal to the ratio of the sum of the diagonal elements to the sum of the elements of the whole matrix, which is equivalent to the accuracy. We multiplied the sum of the elements of row i and the sum of the elements of column i and then added them. Then, we divided the sum by the square of the sum of all elements in the matrix; the resulting ratio was u_a.

3.3. Dataset Statistics

The dataset FORSD, which has both stance labels and true or false labels, is based on the original public fake news dataset. The statistics of the stance labels are shown in Figure 1. FORSD describes the consistency between news headlines and news content through four categories of stance labels: "Agree", "Disagree", "Discusses", and "Unrelated". FORSD contains 2100 news documents, of which the news judged as "discussed" accounts for 50% of the entire dataset, while the news judged as "agree" accounts for about 30%. The rest, judged as "unrelated" and "disagree" news, accounts for about 13% and 7%.

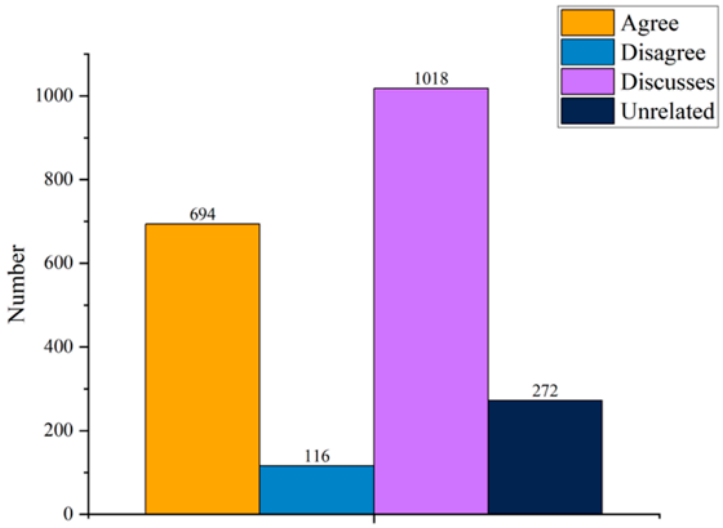

Figure 1. FORSD dataset statistics.

The distribution of FORSD's annotation results generally conforms to the fake news stance expression characteristics of the news frame theory. In order to arouse readers' interest in reading the news text, news writers often express clear opinions or give positive conclusions in the headline. News with high dissemination effectiveness tends to gain a high degree of trust in the news content from readers. One type of news that achieves high trust is characterized by the fact that the headline and the conclusions and opinions in the main body are consistent with each other. Meanwhile, vague discussions of the authenticity of events provide only ambiguous opinions or conclusions. In FORSD, the proportion of news marked as "discussed" and "agree" in the entire dataset is much higher than that of news marked as "disagree" and "unrelated", which proves the relationship between news characteristics and news dissemination. It is an example of the FORSD's data, as presented in Table 2.

Table 2. An example of the FORSD dataset.

Title 1	Text	Label	Stance
You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fellow at the Freedom Center, is a New York writer focusing on radical Islam. In the final stretch of the election, Hillary Rodham Clinton has gone to war with the FBI.	Fake	Agree
Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Monday that he will stop in Paris later this week, amid criticism that no top American officials attended Sunday's unity march against terrorism.	Real	Disagree
Tehran, USA	I'm not an immigrant, but my grandparents are. More than 50 years ago, they arrived in New York City from Iran. I grew up mainly in central New Jersey, an American kid playing little league for the Raritan Red Sox and soccer for the Raritan	Fake	Unrelated
Donald Groped Hillary in 2005! Trump and Weiner Sext Each Other!	Topics: anthony weiner, presidential politics, American Politics, Donald J. Trump, Groping, Clinton's emails Friday, 4 November 2016	Fake	Discuss

4. Problem Statement

Our goal was to build an explainable fake news analysis model which introduces stance information. Given a piece of news data with text content N, set the title T as the target. A represents the stance of the title, and B represents the stance of the body; we need to compare A and B.

Stance Classification: If A and B are consistent, the news internal stance is considered to be classified as agree; if A and B are inconsistent, the news internal stance is considered to be classified as disagree; if A or B does not involve any stance information, the news internal stance is classified as unrelated; if the relationship between A and B cannot be determined, the news internal stance is classified as discuss. Therefore, it can be determined that the mapping relationship between the two is:

$$D_T : N \rightarrow \{\text{Agree, Disagree, Discuss, Unrelated}\} \forall n \in N$$

Explainable Fake News Analysis: According to the propensity score matching method in causal inference, explainable news analysis is defined as a propensity value matching calculation problem between stance information and news classification, and its purpose is to determine the correlation between stance information and news classification. Formally, stance information aims to find a relationship:

$$D_T : N \rightarrow Y$$

where $Y \in \{0, 1\}$ represent the fake news and true news.

5. Method

5.1. Overall Framework

The overall architecture is shown in Figure 2. We introduced a novel explainable fake news analysis method that enhances the explanation and credibility of fake news analysis by combining news classification and stance information. Our model comprises the following components:

(1) Integrated Data Augmentation Module: Aiming at the imbalance of stance classification data in the dataset, we proposed an integrated data augmentation algorithm. This module primarily conducts data augmentation on existing data to form an extended dataset.

(2) Stance Classification Module: We introduced a stance classification model based on a pre-trained model. Specifically, the BERT pre-trained model is used for text representation, followed by the classification task.

(3) Explainable Analysis Module: In order to make full use of the stance information of news, an explainable analysis module that combines stance information was proposed. Here, we employed the propensity score matching method to analyze the relationship between stance consistency and fake news, thereby providing interpretable reasons for the classification of fake news.

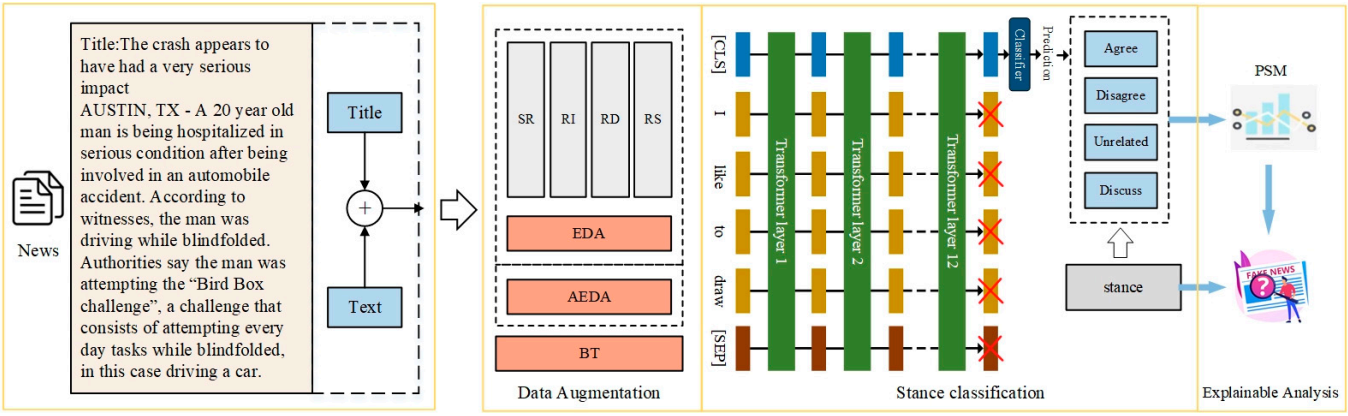

Figure 2. Illustration of the explainable fake news analysis with the stance information framework.

In the SC-FNA (Stance Classification Fake News Analysis) process, given a piece of news data, the following steps are undertaken. The data augmentation module is utilized to form an extended dataset for research. Then, the data are input into the stance classification module, which is based on the BERT pre-trained model, and the stance semantic features are extracted for stance classification. Subsequently, the news data with stance classifications undergoes analysis through the fake news analysis module. In this step, the propensity score matching calculation method is employed, utilizing the stance classification information to explain the relationship between the stance consistency and the true or false classification of the news.

5.2. Data Augmentation Network

EDA (easy data augmentation) [32] is a data expansion method to generate new samples (add-data) from the original text. It involves utilizing the training set to build the model that improves classification performance and generalization ability. EDA has four basic operations: random insertion (RI), random deletion (RD), random swap (RS), and synonym replacement (SR).

AEDA (an easy data augmentation) [33] was proposed as a method for achieving data augmentation by randomly inserting punctuation marks. The approach involves randomly choosing the length of a sequence of numbers between 1/3 and 1 to represent random multiple insertions. The positions in the sequence are also randomly assigned as many as the number of randomly chosen punctuation marks. Finally, for each position selected, a punctuation mark is randomly selected from the six punctuation marks {".", ";", "?", ":", "!", ","} for punctuation insertion. An example of AEDA is shown in Table 3.

Table 3. An example of AEDA.

AEDA	Sentence
NONE	On top of that, Papadopoulos wasn't just a covfefe boy for Trump
AEDA.eg1	On top of. that, Papadopoulos wasn't just a covfefe boy; for Trump;
AEDA.eg2	On, top., of that, Papadopoulos wasn't just a covfefe boy; for Trump
AEDA.eg3	: On, top of. that, Papadopoulos wasn't just a covfefe boy; for Trump!

BT (back-translation) is a data augmentation strategy which has been proved to be effective and stable in previous research experiments. The specific technical methods are as follows: (a) perform a translation operation on the sentence x and translate the source language L1 into the intermediate language L2; (b) retranslate the intermediate language L2 back into the source language L1 to obtain the text data after data enhancement. Among them, the language set L includes multiple languages, such as simplified Chinese, English, French, Spanish, and other intermediate languages.

The EDA, AEDA, and BT text data augmentation methods are all common and effective text augmentation methods, but these three methods are independent of each other. Inspired by ensemble learning, we combined three methods in different ways instead of choosing a complex mathematical algorithm. We used a large number of experiments to obtain the best combination of methods and finally obtained extended text to obtain more training data and text features.

Specifically, in the integrated data enhancement module, we used the EDA and AEDA algorithms. In each iteration, a text was input based on the parameter "naug", and five operation functions (RS, RI, RD, SR, insert punctuation) were performed to create augmented data. For the BT module integrating the data augmentation algorithm, in each iteration, the source language of the text to be translated was set to English, the translation language was set to Spanish, and the Spanish text was obtained through translation. Then, we input the Spanish text into the back-translation process, set the translation source language to Spanish, and set the translation language to English to obtain the output result of the back-translation module. The overall frame diagram is shown in Figure 3.

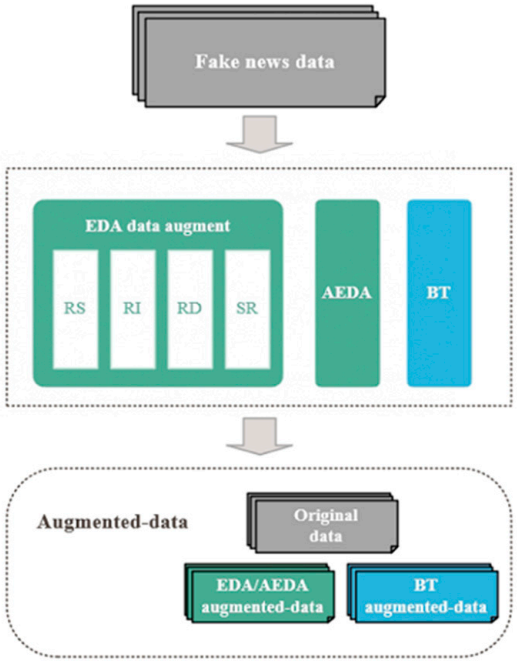

Figure 3. Integrated Data Augmentation Module.

5.3. A Model for Stance Classification

We used the BERT-based [34] pretrained model as a classifier for stance classification. The main pre-trained process of the BERT pre-trained model is to simultaneously train a masked language model using large-scale corpus data and make the following sentence prediction. When the entire pre-trained process is complete, the BERT pre-trained model can be used for downstream tasks. The overall frame diagram is shown in Figure 4.

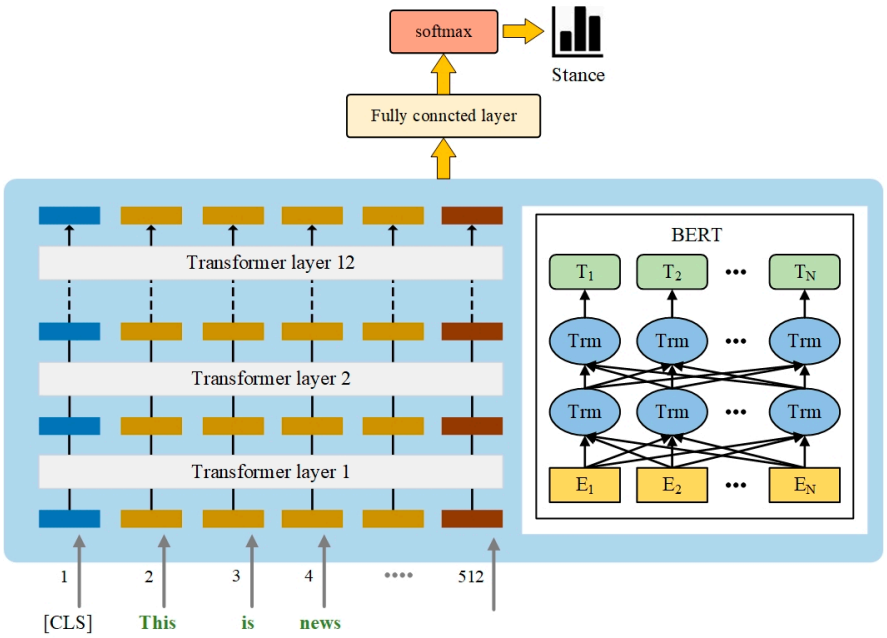

Figure 4. Stance Classification Module.

- Input layer

For a sentence $x_1x_2 \ldots x_n$, y is the original input of BERT. x is used for input. x is obtained by the correlation mapping of y. The relevant mapping contents are: (1) word vector matrix; (2) block vector matrix; (3) position vector matrix. The specific formula is:

$$y = [CLS]x_1x_2 \ldots x_n[SEP] \tag{2}$$

$$x = IR(X) \tag{3}$$

where n is the sentence length. $[CLS]$ is the special marker for the start of text sequences. $[SEP]$ is the separation marker between text sequences.

- BERT encoding layer

The BERT coding layer consists of a 12-layer Transformer encoder structure. In the coding layer, the input representation x is encoded by a multi-layer encoder, and the self-attention mechanism in the model structure can be used to perform the semantic association of words and then obtain the sentence with context. The semantic representation of relation is $l \in R^{N \times o}$. The symbol o is used to denote the hidden layer dimension of BERT.

$$l = BERT(x) \tag{4}$$

Given that the NSP task is performed in the BERT pre-training stage, in which BERT uses $[CLS]$ bit prediction, the same method is usually chosen for classification prediction in text classification tasks. Specifically, l_0 is represented by the hidden layer, which corresponds to the $[CLS]$ bit corresponding to the model; in the specific model structure, $[CLS]$ is the first element in the sequence value input, so l_0 value is constructed from the representation of the first component of l.

- Classification output layer

After the BERT coding layer, the hidden layer representation l_0 corresponding to the $[CLS]$ bit can be obtained, and then a fully connected layer needs to be used for prediction.

Finally, the predicted classification label corresponding to the input text is obtained. The specific calculation formula is:

$$B = softmax\left(l_0 k^0 + n^0\right) \tag{5}$$

where $k^0 \in R^{d \times w}$ represents the weight of the fully connected layer. $n^0 \in R^w$ represents the fully connected layer bias. w represents the number of classification labels. Finally, the probability distribution $B \in R^w$ and the real label value t in the classification are subjected to cross-entropy loss learning, which can realize the learning of model parameters.

5.4. Explainable News Analysis Method

The concept of propensity score [35] originated in 1983. It refers to the conditional probability that a research object is affected by certain independent variables while being able to control for the observed confounding variables. This causal analysis approach is reliable because the researcher can control the propensity score to reduce the effect of selection error on causal conclusions. Further, propensity score matching is a specific method of controlling the propensity score.

Specifically, propensity score matching involves pairing the research objects that can be affected by the independent variable with the research objects that the independent variable does not affect. This ensures that the matched research objects have equal or similar propensity score. Therefore, the basic steps of propensity score matching include three steps:

- Calculate propensity score;
- Matching calculation by propensity score;
- Match the sample values to calculate the causal coefficient.

Using the BERT pre-trained model allows for the conversion of high-dimensional and massive data into structured data, providing a basis for subsequent causal analysis.

In the fake news explainable analysis module, propensity score matching (PSM) is used to infer the causal relationship, that is, the causal relationship between the consistency stance of news headlines and body text and the authenticity of news. Based on the mainstream media's news text framing theory, we proposed the following research causal hypotheses:

- If the stance of news headlines is inconsistent with the stance of news body, news is more likely to be fake news.
- If the stance of news headlines is consistent with the stance of news body, news is more likely to be true news.

The FORSD dataset is a dataset with both stance and true or false. The SC-FNA model can be used for stance classification, and the fake_or_real news dataset is used for stance labeling. We can obtain 6335 pieces of data with stance labels, and the ratio of true to false is 1:1. Before performing the propensity score matching calculation, 750 pieces of data were randomly selected. The specific research steps were:

Step 1: Calculate the propensity score

The propensity value $e(L_i)$ represents the probability that a sample i in the data is affected. If such an effect can be specified as a binary variable, then the propensity score refers to the indicator variable that is affected or not. The calculation formula of the propensity score $e(L_i)$ can be obtained, that is, the effective control of the observable covariate L_i is carried out. Based on this, it is set as a binary variable, then the propensity value refers to the affected or unaffected indicator variables. On this basis, it is set as a binary variable, and the propensity score refers to the indicator variable that is affected or not affected. Finally, the calculation formula of the propensity score $e(L_i)$ is obtained. We performed an

effective control of the observed covariate L_i, and on this basis, we can obtain the probability that the sample i is affected ($M_i = 1$):

$$e(L_i) = P(M_i = 1|L_i) \tag{6}$$

The most significant feature of the propensity score is that it allows for dimensional reduction when there are multiple covariates. It simplifies the dimensionality of the multidimensional covariates to a one-dimensional probability value, which helps balance the covariate between different groups.

Step 2: Match score and effect estimation

The concept of matching refers to the pairing of samples from the experimental group and the comparison group, and the matched samples are samples from different groups but with similar propensity values. In the matching process, if one-to-one strict matching will result in a very small number of available samples, and some samples cannot be matched at all, then an applicable matching method is sought. However, no matter which matching method is used, the calculation expression is the same. The following formula is the utility value of the experimental group after matching the sample:

$$\hat{\alpha}_{TT,M} = \frac{1}{n^1}\sum_i \left[(a_i|b_i = 1) - \sum_j w_{i,j}(a_i|b_i = 0)\right] \tag{7}$$

where n^1 represents the sample size of the experimental group; i and j represent the index values of the experimental group and the comparison sample sequence, respectively; and $w_{i,j}$ represents the weight of the compared samples when the samples are repeatedly matched. The calculation formula of $\hat{\alpha}_{TT,M}$ and ATT (Average Treatment Effect on the Treated) is:

$$\begin{aligned} ATT = E\left(a^1 - a^0 \middle| b = 1\right) &= E\left(a^1 \middle| b = 1\right) - \underbrace{E\left(a^0 \middle| b = 1\right)}_{\text{can't observe}} \\ &= E\left(a^1 \middle| b = 1, e(L)\right) - \underbrace{E\left(a^0 \middle| b = 0, e(L)\right)}_{match\ the\ sample\ of\ control\ group} \end{aligned} \tag{8}$$

In this study, the samples of the experimental group could not observe the counterfactual result, that is, $E\left(a^0 \middle| b = 1\right)$. Therefore, we used the propensity score matching method to explore the research hypothesis at the same time, choosing $E\left(a^0 \middle| b = 0\right)$ instead of $E\left(a^0 \middle| b = 1\right)$.

5.5. Training

The formula for calculating the cross-entropy loss is as follows.

$$Loss = -\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{c} y_j^{(i)} log\hat{y}_j^{(i)} \tag{9}$$

where $y_j^{(i)}$ represents the actual output result, and $\hat{y}_j^{(i)}$ represents the predicted probability of the model. i is used to represent the sample, and j is used to represent the category.

When there is a significant discrepancy between the expected value and the actual value, it indicates that the model's prediction performance is poor, and the negative logarithm is infinite; on the contrary, when the predicted value and the actual value are close, the prediction performance of the model is good. According to the formula, the logarithm takes a negative value, and it can be seen that the entire change law is exponential, which can be summarized as follows: the model performs poorly, the loss function gradient is large, and the model learns quickly; the model performs better, the loss function gradient is small, and the model learns slowly.

According to the above analysis, the essence of the cross-entropy loss function is to operate the Bernoulli distribution of the multi-category output results and finally realize the maximization of the log-likelihood function. For the stance detection task, stance is defined as four categories of "agree", "disagree", "irrelevant", and "discuss", which is a typical text multi-classification task. Therefore, the cross-entropy loss function was used in this experiment. A further simplification, the cross-entropy loss function for the negative log-likelihood loss is expressed as follows:

$$Loss = -\frac{1}{n}\sum_{i=1}^{n} log\hat{y}_m^{(i)} \tag{10}$$

where $\hat{y}_m^{(i)}$ represents the prediction probability of the model for the sample on the correct class m.

6. Experiments and Results

6.1. Experimental Setup

Step 1 (datasets): The FNC1 dataset is the competition dataset for the Fake News Challenge [36], the statistics of the FNC1 dadaset as shown in Table 4. The FNC1 dataset consists of pairs of title and body texts, each with a corresponding stance classification label. The dataset has four files: train_bodies.csv, consisting of text body and ID; train_stances.csv, consisting of stance classification, article title, and ID; test_bodies.csv, consisting of text body, ID, and position classification; and article test_stances.csv, consisting of title and ID. The train dataset contains 64,205 data entries, while the test dataset contains 28,972 data entries.

Table 4. The statistics of the FNC1 dataset.

Stance	Number	
	Train	**Test**
Agree	4935	2237
Disagree	2242	1069
Unrelated	9813	4643
Discusses	47,215	21,023
total	64,205	28,972

Step 2 (evaluation metrics): Precision is the classification accuracy, which is mainly used to measure whether the classifier can correctly identify the category of the sample during the classification process. It is often referred to as the precision rate. If it corresponds to the positive samples and negative samples, precision represents the proportion of samples identified as positive samples during the classification process, among which the proportion of correctly predicted samples. Recall, also known as the recall rate, is another classification accuracy metric that represents the proportion of correctly predicted positive samples among all actual positive samples. The $F1 - score$ is a comprehensive evaluation index that takes into account both precision and recall.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F_b = \frac{\left(1 + b^2\right) \cdot Precision \cdot Recall}{b^2 \times Precision + Recall} \tag{13}$$

Compared with the two-class classification, the evaluation indicators of the multi-classification are relatively complex. In order to more accurately judge the accuracy of each classification in the prediction probability, considering the class imbalance problem of the multi-classification dataset, the weight of each class is calculated so that the precision can be known. Recall is the weighted average of the corresponding precision and recall for each category. The corresponding calculation formula of $Weight - F1$ is:

$$Precision_l = \frac{TP_l}{TP_l + FP_l} \tag{14}$$

$$Precision_{weighted} = \frac{\sum_{l=1}^{L} Precision_l \times w_l}{|L|} \tag{15}$$

$$Recall_l = \frac{TP_l}{TP_l + FN_l} \tag{16}$$

$$Recall_{weighted} = \frac{\sum_{l=1}^{L} Recall_l \times w_l}{|L|} \tag{17}$$

Step 3 (implementation details): The hidden layer dimension is 256, the learning rate is set to 2×10^{-5}, and the dropout rate is 0.1. Len_max is set to 350 for the maximum sequence length.

6.2. Baseline

In order to verify the effectiveness of the BERT pre-trained model used in this paper on the performance of the multi-stance classification task, different classical network models were introduced as comparison models in the experiments. In the experiments, these models were used in the selected datasets at the same time, and their parameters were adjusted based on the specific dataset conditions to optimize the classification accuracy.

- SVM [37]: A Support Vector Machine model performs stance classification relying on manually extracted features.
- LSTM [38]: The most basic LSTM network model was selected as one of the comparison methods. This model was applied to various NLP tasks. There is no attention mechanism in the network structure. This model retains the word order relationship of the input text features but does not have an attention mechanism.
- BIMPM [39]: This model, proposed in the FNC1 Fake News Challenge competition, is similar to checking matching verbatim, using a bidirectional RNN to separately represent the contextual representation of the headline and body text and then computing the cosine similarity between the headline and body context representation vectors.
- featMLP [40]: featMLP is the model that achieved second place in the FNC, which is an ensemble of multi-layer perceptron (MLP) with six hidden layers and a Softmax layer each. This system uses a variety of models to obtain the feature.
- BiLSTM [41]: The Long Short-Term Memory (LSTM) model is an optimization model for recurrent neural networks, which can solve the gradient disappearance problem that arises in long sequence learning. To learn contextual information better, a bidirectional LSTM (BiLSTM) model is proposed based on LSTM. BiLSTM is a combination of a forward LSTM and a backward LSTM, capable of encoding the input data in both directions.
- BiLSTM-Attention [42]: BiLSTM-Attention is a neural network structure that combines BiLSTM and the attention mechanism. BiLSTM consists of bidirectional LSTMs, which encode data with LSTMs in both the forward and reverse directions and integrate and output high-dimensional features. The output of BiLSTM is connected with attention, the features output from each time step are weighted, and the weighted fused feature vector is output.

- BERTbase [34]: The BERT coding layer consists of a 12-layer Transformer encoder structure.
- CNN and DNN with SCM [43]: They use natural language processing technology to process text, reduce extracted features, and use SCM to find similarities between pairs. Then, the new feature is input into the CNN and DNN deep learning methods.
- HeadlineStanceChecker [44]: The position of the title relative to its associated text can be determined. Its novelty lies in the use of a two-level classification architecture, which uses summarization technology to shape the input of the two classifiers rather than directly transmitting the complete news text, thereby reducing the amount of information to be processed while retaining important information. Specifically, the summary is completed through the location language model, and the semantic resources are used to identify the salient information in the text; then, they are compared with the corresponding title.
- ML Ensemble Model [45]: The methodology employed a decentralized Spark cluster to create a stacked ensemble model. Following feature extraction using N-grams, Hashing TF-IDF, and count vectorizer, we used the proposed stacked ensemble classification model.
- Augmentation-based Ensemble Learning [46]: The proposed approach is a mixture of bagging and stacking and leverages text augmentation to enhance the diversity and the performance of base classifiers.
- LightGBM [47]: Light Gradient Boosting Machine (LightGBM) is a popular and efficient machine learning model used for supervised learning tasks, particularly in the domain of gradient boosting.
- SC-FNAours: Our proposed method, which includes a data augmentation algorithm to create an extended dataset and use a pre-trained language model to build a stance classification model for the classification task.

6.3. Results and Analysis

In this section, we conducted a series of experiments to evaluate the performance of the stance classification model using the integrated data augmentation algorithm proposed in this chapter, and to evaluate the role of each parameter and model part, we set up various experiments for comparative analysis. In this section, the public dataset FNC-1 with representative rows was selected for the stance classification experiments, and the self-built dataset FORSD was used to evaluate the performance of the model.

In order to compare the performance of different models in the stance classification tasks and better reflect the advantages of SC-FNA in stance classification, we designed several different sets of comparative experiments: (1) The SVM baseline model was introduced to verify that deep learning was compared with traditional machine learning in stance classification. (2) We verified the performance of multi-classification for different model variants. (3) We introduced the current state-of-the-art stance models in classification tasks, verifying the stance classification performance of the SC-FNA model proposed in this paper.

Table 5 shows the performance of all the compared models based on the two datasets. The following observations can be drawn from the table:

(1) Deep Learning vs. Traditional Machine Learning: The deep learning-based models demonstrated significantly better performance in the stance classification task compared to the traditional machine learning models. This is because the traditional machine learning method uses manual feature extraction, and the information representation ability of these feature extraction methods is relatively poor. In contrast, the model feature extraction ability of deep learning is relatively strong, so it had better performance.

(2) BiLSTM vs. LSTM. The BiLSTM model, which incorporates context information for semantic modeling, outperformed the LSTM model that relies solely on previous in-

formation. This indicates that it is particularly important to understand the information in the context of the news for the judgment of the consistency of the stance.

Table 5. Performance comparison of the proposed SC-FNA method against the baselines.

Model	FNC1		FORSD	
	ACC	**F1**	**ACC**	**F1**
SVM	73.13	63.16	61.26	53.01
LSTM	76.80	69.14	63.59	56.21
BiLSTM	81.29	70.01	70.36	58.23
BiLSTM-attention	82.23	73.21	71.58	61.45
CNN and DNN with SCM	84.60	75.62	73.57	63.85
BiMPM	86.34	84.61	75.93	65.03
LightGBM	87.67	86.48	78.45	69.03
featMLP	88.27	87.08	78.92	70.13
HeadlineStanceChecker	94.31	80.39	81.36	72.45
ML Ensemble Model	93.41	92.40	80.26	78.82
Augmentation-based Ensemble Learning	90.67	90.15	85.78	80.24
$BERT_{base}$	91.32	90.41	86.89	81.18
SC-FNAours	93.85	93.04	86.35	83.60

Indeed, the performance of deep learning in stance classification is generally effective, and adding attention layers to related models can lead to further improvements in performance. The reason is that the addition of the attention layer can effectively capture key information, enhance the model's understanding of text information, and then improve the accuracy of stance classification.

The SC-FNA model proposed in this paper has the best performance and achieved very efficient stance classification performance on the public datasets. There are two main reasons: (1) The algorithm of integrated data augmentation was introduced to expand the dataset, which improved the performance of the model; (2) Using a pre-trained language model for text representation can better capture text information and effectively improve the performance of stance classification tasks.

6.4. Model Ablation

The proposed SC-FNA consists of three components: the data augmentation module, the stance classification module, and the explainable news analysis module. To evaluate the effectiveness of the data augmentation and stance classification module in our method, we ablated our method into several simplified models and compared their performance with related methods. The details of these methods are described as follows:

SC-FNAbase: The integrated data enhancement module was removed from the SC-FNA model, and only stance classification was performed through $BERT_{base}$.

SC-FNA-EDA: Only the EDA data augmentation algorithm with n_{aug} set to 16 was used in the integrated data augmentation module from the SC-FNA model.

SC-FNA-AEDA: Only the AEDA data augmentation algorithm with n_{aug} set to 16 was used in the integrated data augmentation module from the SC-FNA model.

Non-BT SC-FNA: From the SC-FNA model, the EDA and AEDA data enhancement algorithms with n_{aug} set to 16 were used in the integrated data augmentation module, and the BT data augmentation algorithm was removed.

SC-FNA-BT: Only the BT data augmentation algorithm with n_{aug} set to 1 was used in the integrated data augmentation module from the SC-FNA model.

Non-AEDA SC-FNA: From the SC-FNA model, only the BT with n_{aug} set to 1 and the EDA data augmentation algorithm with n_{aug} set to 16 were used in the integrated data augmentation module, and the AEDA data augmentation algorithm was removed.

Non-EDA SC-FNA: The BT with n_{aug} set to 1 and the AEDA data augmentation algorithm with n_{aug} set to 16 were used in the integrated data augmentation module, and the EDA data augmentation algorithm was removed.

As shown in Table 6, we compared the stance classification performance of the SC-FNA variants on FNC-1 and FORSD dataset. In Table 7, we show the stance detection results of the SC-FNA variants on the FORSD dataset.

Table 6. Comparison among SC-FNA variants (P: PRECISION; R: RECALL; F1: F1 SCORE).

Model	FNC1			FORSD		
	P	R	F1	P	R	F1
SC-FNAbase	90.65	90.17	90.41	77.4	85.34	81.18
S-FND-EDA	91.45	89.81	90.62	81.45	81.70	81.57
SC-FNA-AEDA	92.45	90.68	91.56	79.56	84.17	81.80
Non-BT SC-FNA	90.24	93.66	91.92	80.12	84.22	82.12
SC-FNA-BT	91.90	91.92	91.91	79.95	84.16	82.00
Non-AEDA SC-FNA	91.40	92.91	92.15	81.23	83.48	82.34
Non-EDA SC-FNA	90.23	95.24	92.67	82.18	82.82	82.50
SC-FNA	91.59	94.53	93.04	82.59	84.63	83.60

Table 7. The stance classification performance (F1) comparison among SC-FNA variants on FORSD.

Model	Agree	Disagree	Unrelated	Discusses
SC-FNAbase	80.23	45.56	60.12	78.56
S-FND-EDA	80.56	46.67	60.18	78.79
SC-FNA-AEDA	80.79	47.59	61.45	79.12
Non-BT SC-FNA	81.13	48.79	61.85	79.48
SC-FNA-BT	81.59	49.16	62.48	79.98
Non-AEDA SC-FNA	82.01	50.21	62.59	80.12
Non-EDA SC-FNA	81.98	50.79	63.71	80.45
SC-FNA	82.36	50.86	64.58	80.47

From the presented tables, it is evident that the SC-FNA model achieves good performance, which indicates that the integrated data augmentation algorithm can be an important supplementary means in fake news datasets with imbalanced classification. From the Table 7, we can find that all the modules proposed in our paper can benefit from the stance multi-classification task. Based on the results from these tables, we can conclude that:

After applying the SC-FNA model and the integrated data augmentation module, the outcome provides evidence that the data augmentation algorithm indeed enhanced the model's generalization ability by expanding the dataset.

At the same time, the performance improvement of SC-FNA-EDA compared with SC-FNAbase was only 0.21%, indicating that the simple data augmentation algorithm has little effect on the performance of stance classification. After adding the AEDA data augmentation algorithm, the performance of Non-BT SC-FNA was improved by 1.51%, which represents a significant improvement.

The performance of SC-FNA was 1.12% higher than that of Non-BT SC-FNA, which proves that the back translation data augmentation algorithm effectively improves the model's performance and contributes to the enhanced accuracy of stance classification.

6.5. Sensitivity Analysis

We demonstrated the effectiveness of the data augmentation module in improving model performance through ablation experiments. In this subsection, we evaluated the effect of the parameter naug. We first set the naug to the [4,8,16,32] dimension for g, then optimized the rest of the hyperparameters on the validation subset. In Figure 5, we illustrate the effect of varying naug values on the model's performance for both datasets. We observe that the results were similar for both evaluation measures, Accuracy and F1. Setting the data to 16 brought a larger performance improvement to the model. When the data were 16, the model obtained the highest F1 of 90.62% on the FNC-1 test subset and the highest F1 of 81.57% on the FORSD test subset. These results show that maintaining an increase in the data of n can lead to an improvement in performance. Setting the data from 16 to 32 led to performance degradation in the model. This suggests that when the model becomes too complex, it may suffer from overfitting issues, resulting in reduced performance.

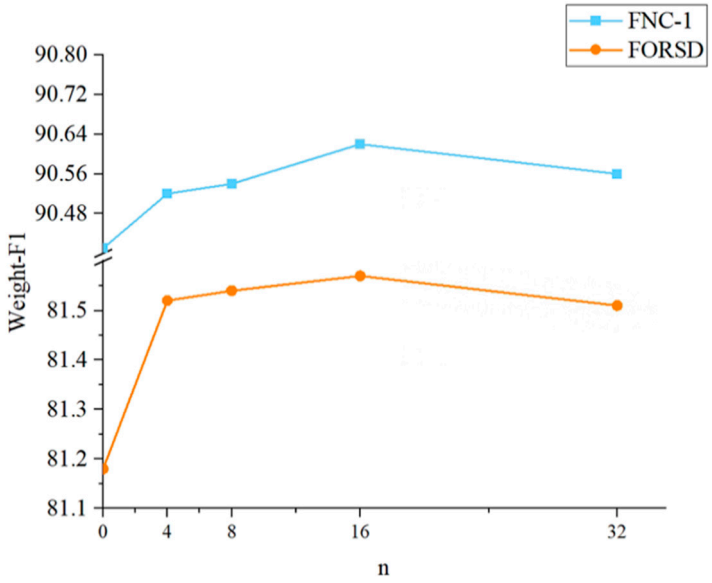

Figure 5. The effect of the data augmentation parameter on model performance.

6.6. Explainable Analysis

In the design of the explainable analysis module, propensity score matching was used to explore the impact of stance consistency on the formation of fake news. To achieve this, we conducted an analysis of the confounding variables present in the news data and then calculated the tendency value to analyze the relationship between stance consistency and fake news.

Confounding variable analysis of news data: In this part, we analyzed the intuitive impact of real news and fake news on the main factors in the news framework theory, including the length of news content (divided into long news, short news), the length of news headlines (short headlines, long headlines), and news categories (technology, society, health). We attempted to control these elements to form control variables.

The impact of stance consistency on the fake news: In this study, we utilized the logit model to estimate the propensity score of the stance. We then performed a detection analysis of the matching quality of the propensity score of the stance consistency. The experimental data indicated that the logit regression results were effective, suggesting a good degree of fit. Furthermore, the variables between the matched experimental group and the control group were balanced. Thus, the supporting hypothesis on confounding variable

control of propensity score matching was satisfied. Various matching methods were employed in this study, including the OLS, nearest neighbor matching, kernel and local linear matching, and radius matching. After performing matching based on the propensity score, we used true news and fake news as the dependent variables to test the impact of position consistency on fake news. The results of these analyses are presented in Table 8 below.

Table 8. The effect of stance classification on news authenticity. (*, **, *** Indicate significance at the 0.10, 0.05, and 0.01 levels).

	OLS		**Nearest Neighbor Matching**		**Kernel and Local Linear Matching**		**Radius Matching**	
	ATT	**T**	**ATT**	**T**	**ATT**	**T**	**ATT**	**T**
Real News	0.192 *	1.91	0.245 *	1.72	0.245 *	2.25	0.31 ***	2.63
Fake News	−1.851 ***	−2.8	−2.213 **	−2.53	−1.531 **	−2.75	−1.312 **	−2.32

The experimental data show that stance consistency can significantly improve the possibility of true news, and the results of OLS and the other three tendencies matching can also support this conclusion through data, so the correlation between stance consistency and true news is positive and robust. At the same time, the relationship between stance consistency and fake news is negative. It can be inferred that fake news is more inclined to exhibit inconsistent stances, and this result was also cross-validated by the four matching methods.

In conclusion, based on the FORSD dataset, we used propensity score matching to calculate the relationship between stance consistency and fake news. Our findings indicate a strong correlation between a consistent stance and real news, as well as a strong correlation between an inconsistent stance and fake news. The results show that there is a correlation between the consistent expression of stance in headlines and body text in news articles and the authenticity of fake news. This study provides a research premise and evidence support for the explainable analysis of fake news.

7. Conclusions

In this paper, we proposed a fake news analysis method called Stance Classification for Fake News Analysis (SC-FNA), which aims to leverage cognitive and information science principles to foster the sustainable use of technology and achieve a more explainable artificial intelligence system. Specifically, we aimed to introduce stance information to improve the credibility of fake news analysis. Based on existing public datasets of fake news, we used cognitive surveys to exclude partisan bias and then annotated the classification of stance information to form a dataset that could be used for explainable fake news analysis research. The integrated data enhancement algorithm effectively solves the problem of imbalanced data classification. We used the propensity score matching method for causal inference to verify the correlation between stance consistency and news authenticity, making the fake news analysis results more explainable. Extensive experiments on the public datasets demonstrate that our proposed method achieves an effective performance.

Indeed, this study has certain limitations that should be acknowledged. First, our analysis was primarily focused on the text aspect of fake news, neglecting the potential impact of audio and video content in the propagation of misinformation. Future studies should further explore the multimodal fake news dataset. Second, considering the incorporation of multimodal datasets of large models and the emergence of AIGC, the need for explainable methods becomes paramount.

Author Contributions: All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by L.Y. The first draft of the manuscript was written by L.Y. Software, H.S.; Data curation, H.J.; Writing—review and editing, N.C. and L.S. All authors commented on previous versions of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China under Grant No. 2020YFF0305300 and No. 2022YFC3302103, the Fundamental Research Funds for the Central Universities (No. CUC23GY005).

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nasir, J.A.; Khan, O.S.; Varlamis, I. Fake news detection: A hybrid CNN-RNN based deep learning approach. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100007. [CrossRef]
2. Kaliyar, R.K.; Goswami, A.; Narang, P.; Sinha, S. FNDNet—A deep convolutional neural network for fake news detection. *Cogn. Syst. Res.* **2020**, *61*, 32–44. [CrossRef]
3. Capuano, N.; Fenza, G.; Loia, V.; Nota, F.D. Content-Based Fake News Detection with Machine and Deep Learning: A Systematic Review. *Neurocomputing* **2023**, *530*, 91–103. [CrossRef]
4. Seddari, N.; Derhab, A.; Belaoued, M.; Halboob, W.; Al-Muhtadi, J.; Bouras, A. A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media. *IEEE Access* **2022**, *10*, 62097–62109. [CrossRef]
5. 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, Melbourne, VIC, Australia, 11–15 February 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 312–320.
6. Monti, F.; Frasca, F.; Eynard, D.; Mannion, D.; Bronstein, M.M. Fake News Detection on Social Media using Geometric Deep Learning. *arXiv* **2019**, arXiv:1902.06673.
7. Mishima, K.; Yamana, H. A Survey on Explainable Fake News Detection. *IEICE Trans. Inf. Syst.* **2022**, *105*, 1249–1257. [CrossRef]
8. Kasnesis, P.; Toumanidis, L.; Patrikakis, C.Z. Combating Fake News with Transformers: A Comparative Analysis of Stance Detection and Subjectivity Analysis. *Information* **2021**, *12*, 409. [CrossRef]
9. Davoudi, M.; Moosavi, M.R.; Sadreddini, M.H. DSS: A hybrid deep model for fake news detection using propagation tree and stance network. *Expert Syst. Appl.* **2022**, *198*, 116635. [CrossRef]
10. Salah, I.; Jouini, K.; Korbaa, O. Augmentation-based ensemble learning for stance and fake news detection. In Proceedings of the International Conference on Computational Collective Intelligence, Hammamet, Tunisia, 28–30 September 2022; Springer International Publishing: Cham, Switzerland, 2022; pp. 29–41.
11. Sengan, S.; Vairavasundaram, S.; Ravi, L.; AlHamad, A.Q.M.; Alkhazaleh, H.A.; Alharbi, M. Fake News Detection Using Stance Extracted Multimodal Fusion-Based Hybrid Neural Network. *IEEE Trans. Comput. Soc. Syst.* **2023**, 1–12. [CrossRef]
12. Küçük, D.; Can, F. Stance Detection: Concepts, Approaches, Resources, and Outstanding Issues. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 11–15 July 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 2673–2676.
13. Küçük, D.; Can, F. Stance detection: A survey. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–37. [CrossRef]
14. Dey, K.; Shrivastava, R.; Kaushik, S. Topical Stance Detection for Twitter: A Two-Phase LSTM Model Using Attention. *arXiv* **2018**, arXiv:1801.03032.
15. Hanselowski, A.; Pvs, A.; Schiller, B.; Caspelherr, F.; Chaudhuri, D.; Meyer, C.M.; Gurevych, I. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1859–1874.
16. Zhang, Q.; Yilmaz, E.; Liang, S. Ranking-based Method for News Stance Detection. In Proceedings of the Web Conference 2018, Lyon, France, 23–27 April 2018; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2018; pp. 41–42.
17. Ghanem, B.; Rosso, P.; Rangel, F. Stance Detection in Fake News a Combined Feature Representation. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Brussels, Belgium, 1 November 2018; pp. 66–71.
18. Li, Y.; Song, Y. A survey of text stance detection. *J. Comput. Res. Dev.* **2021**, *58*, 2538–2557.
19. Lai, M.; Cignarella, A.T.; Farías, D.I.H.; Bosco, C.; Patti, V.; Rosso, P. Multilingual stance detection in social media political debates. *Comput. Speech Lang.* **2020**, *63*, 101075. [CrossRef]
20. Vychegzhanin, S.; Kotelnikov, E. Stance Detection in Russian: A Feature Selection and Machine Learning Based Approach. *AIST (Suppl.)* **2017**, *12*, 166–177.
21. Aldayel, A.; Magdy, W. Stance detection on social media: State of the art and trends. *Inf. Process. Manag.* **2021**, *58*, 102597. [CrossRef]

22. Alturayeif, N.; Luqman, H.; Ahmed, M. A systematic review of machine learning techniques for stance detection and its applications. *Neural Comput. Appl.* **2023**, *35*, 5113–5144. [CrossRef]
23. Küçük, D.; Can, F. A tutorial on stance detection. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Virtual, 21–25 February 2022; pp. 1626–1628.
24. Zhang, B.; Ding, D.; Jing, L. How would stance detection techniques evolve after the launch of chatgpt? *arXiv* **2022**, arXiv:2212.14548.
25. Chien, S.-Y.; Yang, C.-J.; Yu, F. XFlag: Explainable Fake News Detection Model on Social Media. *Int. J. Hum. Comput. Interact.* **2022**, *38*, 1808–1827. [CrossRef]
26. Wu, K.; Yuan, X.; Ning, Y. Incorporating Relational Knowledge in Explainable Fake News Detection. In *Advances in Knowledge Discovery and Data Mining*; Springer: Cham, Switzerland, 2021; pp. 403–415.
27. Qiao, Y.; Wiechmann, D.; Kerz, E. A Language-Based Approach to Fake News Detection Through Interpretable Features and BRNN. In Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM), Barcelona, Spain, 13 December 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 14–31.
28. Lu, Y.-J.; Li, C.-T. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; pp. 505–514.
29. Chi, H.; Liao, B. A quantitative argumentation-based Automated eXplainable Decision System for fake news detection on social media. *Knowl.-Based Syst.* **2022**, *242*, 108378. [CrossRef]
30. Ni, S.; Li, J.; Kao, H.-Y. MVAN: Multi-View Attention Networks for Fake News Detection on Social Media. *IEEE Access* **2021**, *9*, 106907–106917. [CrossRef]
31. Luo, Y.; Card, D.; Jurafsky, D. Detecting Stance in Media on Global Warming. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Virtual, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 3296–3315.
32. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 6382–6388.
33. Karimi, A.; Rossi, L.; Prati, A. AEDA: An Easier Data Augmentation Technique for Text Classification. In Proceedings of the 8th Workshop on Argument Mining, Punta Cana, Dominican Republic, 10–11 November 2021; pp. 2748–2754.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
35. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
36. Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; Zubiaga, A. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In Proceedings of the 11th International Workshop Semantic Evaluation, Vancouver, BC, Canada, 3–4 August 2017; pp. 69–76.
37. Krishna, N.L.S.R.; Adimoolam, M. Fake News Detection system using Decision Tree algorithm and compare textual property with Support Vector Machine algorithm. In Proceedings of the 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 16–17 February 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
38. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, (Volume 2: Short Papers), Calgary, AB, Canada, 17–22 August 2020; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 207–212.
39. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
40. Zeng, Q.; Zhou, Q.; Xu, S. Neural Stance Detectors for Fake News Challenge. *CS224n: Natural Language Processing with Deep Learning*. 2017. Available online: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761936.pdf (accessed on 4 June 2023).
41. Slovikovskaya, V. Transfer learning from transformers to fake news challenge stance detection (FNC-1) task. *arXiv* **2019**, arXiv:1910.14353.
42. Chen, T.; Xu, R.; He, Y.; Wang, X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Syst. Appl.* **2017**, *72*, 221–230. [CrossRef]
43. Jawad, Z.A.; Obaid, A.J. Combination of Convolution Neural Networks And Deep Neural Networks For Fake News Detection. *arXiv* **2022**, arXiv:2210.08331.
44. Sepúlveda-Torres, R.; Vicente, M.; Saquete, E.; Lloret, E.; Palomar, M. HeadlineStanceChecker: Exploiting summarization to detect headline disinformation. *J. Web Semant.* **2021**, *71*, 100660. [CrossRef]
45. Altheneyan, A.; Alhadlaq, A. Big Data ML-Based Fake News Detection Using Distributed Learning. *IEEE Access* **2023**, *11*, 29447–29463. [CrossRef]

46. Salah, I.; Jouini, K.; Korbaa, O. On the use of text augmentation for stance and fake news detection. *J. Inf. Telecommun.* **2023**, *7*, 359–375. [CrossRef]
47. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.