algorithms

MDPI

Article

MVACLNet: A Multimodal Virtual Augmentation Contrastive Learning Network for Rumor Detection

Xin Liu[1,*], Mingjiang Pang [1], Qiang Li [2], Jiehan Zhou [3], Haiwen Wang [1] and Dawei Yang [1]

[1] Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), No. 66, West Changjiang Road, Huangdao District, Qingdao 266580, China; sunnypmj@163.com (M.P.); w824569950@163.com (H.W.); 18351805913@163.com (D.Y.)
[2] Scientific and Technological Innovation Center of ARI, Beijing 100020, China; liqiang@qaii.ac.cn
[3] Information Technology and Electrical Engineering, University of Oulu, 90570 Oulu, Finland; jiehan.zhou@ieee.org
* Correspondence: lx@upc.edu.cn

Abstract: In today's digital era, rumors spreading on social media threaten societal stability and individuals' daily lives, especially multimodal rumors. Hence, there is an urgent need for effective multimodal rumor detection methods. However, existing approaches often overlook the insufficient diversity of multimodal samples in feature space and hidden similarities and differences among multimodal samples. To address such challenges, we propose MVACLNet, a Multimodal Virtual Augmentation Contrastive Learning Network. In MVACLNet, we first design a Hierarchical Textual Feature Extraction (HTFE) module to extract comprehensive textual features from multiple perspectives. Then, we fuse the textual and visual features using a modified cross-attention mechanism, which operates from different perspectives at the feature value level, to obtain authentic multimodal feature representations. Following this, we devise a Virtual Augmentation Contrastive Learning (VACL) module as an auxiliary training module. It leverages ground-truth labels and extra-generated virtual multimodal feature representations to enhance contrastive learning, thus helping capture more crucial similarities and differences among multimodal samples. Meanwhile, it performs a Kullback–Leibler (KL) divergence constraint between predicted probability distributions of the virtual multimodal feature representations and their corresponding virtual labels to help extract more content-invariant multimodal features. Finally, the authentic multimodal feature representations are input into a rumor classifier for detection. Experiments on two real-world datasets demonstrate the effectiveness and superiority of MVACLNet on multimodal rumor detection.

Keywords: rumor detection; multimodal learning; data augmentation; contrastive learning; social media analysis

check for updates

Citation: Liu, X.; Pang, M.; Li, Q.; Zhou, J.; Wang, H.; Yang, D. MVACLNet: A Multimodal Virtual Augmentation Contrastive Learning Network for Rumor Detection. *Algorithms* **2024**, *17*, 199. https://doi.org/10.3390/a17050199

Academic Editor: Ioannis G. Tsoulos

Received: 9 April 2024
Revised: 30 April 2024
Accepted: 1 May 2024
Published: 8 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, with the rapid development of the Internet and intelligent terminals, social media has become an indispensable communication channel in people's daily lives due to its convenience, real-time nature, information sharing ability, interactivity, and diverse content. Although social media facilitates information exchange among people, it has also become a new breeding ground for spreading rumors. Since the outbreak of the COVID-19 epidemic, Facebook and Instagram have deleted over 20 million rumors in total, while Sina Weibo handled 66,251 rumors in 2021 alone. Furthermore, the low cost of information dissemination and imperfect supervision systems also lead to the rapid spread of numerous rumors on social media [1], posing significant threats to individuals' daily lives [2], social stability, and even national security. Therefore, how to effectively and accurately detect rumors on social media has become one of the crucial issues in both academia and industry.

With the development of multimedia technology, rumors spreading on social media have gradually changed from textual forms to multimodal forms with images. Multimodal rumors with images can often create an immersive reading experience to attract readers and give the illusion of high reliability to mislead readers, making it easier to rapidly spread the rumor. However, traditional rumor detection methods [3–16] primarily utilize features extracted from text to detect rumors, which limits their ability to detect such multimodal rumors. Furthermore, previous studies [17–19] have demonstrated that image content in posts effectively serves as supplementary information for rumor detection, thereby improving detection performance. For these reasons, an increasing number of researchers [19–30] have turned their attention to multimodal rumor detection to mitigate the adverse impact of multimodal rumors.

In the real world, aside from rumors unintentionally created and spread by regular users, there are also malicious rumors that are carefully crafted and deliberately spread by rumor creators. Whether through content tampering or relationship network manipulation, the intention of these rumor creators is to make rumors appear indistinguishable from non-rumors in feature space, thus confusing detection models. Moreover, when humans assess the veracity of a post, they generally attempt to find similarities and differences between the post and other confirmed posts to help identify rumors. These phenomena make it vital to consider enhancing the diversity of multimodal samples in feature space and capturing hidden similarities and differences among multimodal samples to assist rumor detection models in adapting to a broader range of complex data distributions and understanding underlying data structures and patterns more effectively for further improvement of their detection performance. Unfortunately, most existing multimodal models tend to overlook these crucial considerations.

Based on the above considerations and to explore a more effective method for multimodal rumor detection, we propose MVACLNet, a Multimodal Virtual Augmentation Contrastive Learning Network. In MVACLNet, we first design a Hierarchical Textual Feature Extraction (HTFE) module to extract textual features from multiple perspectives, namely local, global continuous, and global non-continuous, to achieve a more comprehensive utilization of text data. Then, we enhance the fusion of textual and visual features using a modified cross-attention mechanism, which operates from textual–visual and visual–textual perspectives at the feature value level, to obtain richer and more precise authentic multimodal feature representations. Following this, we devise a Virtual Augmentation Contrastive Learning (VACL) module as an auxiliary training module. This module serves three purposes. One is to generate extra virtual multimodal feature representations and corresponding virtual labels using an interpolation-based data augmentation strategy, regarded as augmented samples, to enhance the diversity of multimodal samples in feature space, thus strengthening the feature learning of the model. The second is to introduce contrastive learning and additionally leverage ground-truth labels and virtual multimodal feature representations to enhance it, thereby helping the model capture more crucial similarities and differences among multimodal samples. The third is to perform the Kullback–Leibler (KL) divergence constraint between predicted probability distributions of the virtual multimodal feature representations and their corresponding virtual labels to help the model capture more content-invariant multimodal features. In particular, the introduced ground-truth labels can provide effective supervisory signals for contrastive learning to prevent the model from misclassifying posts with noise. Finally, a Rumor Classification module is developed to perform rumor prediction on the authentic multimodal feature representations. Overall, with the help of VACL, MVACLNet can learn more robust and generalized multimodal feature representations, thereby improving its detection performance.

In summary, the contributions of our paper are as follows:

- We propose MVACLNet, a Multimodal Virtual Augmentation Contrastive Learning Network, which achieves more effective multimodal rumor detection. It consists of five modules: a Hierarchical Textual Feature Extraction module, a Visual Feature

Extraction module, a Multimodal Feature Fusion module, a Virtual Augmentation Contrastive Learning module, and a Rumor Classification module. Each designed module has a different role, and all the modules contribute to the improvement of detection performance.

- We design a Hierarchical Textual Feature Extraction (HTFE) module to extract textual features from multiple perspectives in order to make comprehensive use of text data.
- We utilize a modified cross-attention mechanism, which operates from different perspectives at the feature value level, to obtain richer and more precise multimodal feature representations.
- We devise a Virtual Augmentation Contrastive Learning (VACL) module as an auxiliary training module to improve detection performance, which can help the model learn more robust and generalized multimodal feature representations by enhancing the diversity of multimodal samples in feature space to enhance feature learning, capturing more crucial similarities and differences among multimodal samples, and extracting more content-invariant multimodal features.
- Experiments on two real-world datasets demonstrate the effectiveness and superiority of MVACLNet in multimodal rumor detection.

2. Related Work

Based on the number of information modalities utilized, existing rumor detection methods can be divided into two main groups: unimodal and multimodal rumor methods.

2.1. Unimodal Rumor Detection Methods

Existing unimodal rumor detection methods utilize features extracted from either texts or images to detect rumors, where the former accounts for the vast majority. Traditional machine learning-based models often rely on hand-crafted textual features to detect rumors [3–6]. For example, Castillo et al. [3] manually constructed more than 80 statistical features, such as punctuation marks, emoticons, and link counts. Zhao et al. [4] designed a set of regular expressions to select signal tweets containing skeptical enquiries. However, these feature engineering-based methods are time-consuming and labor-consuming, which promoted the emergence and rapid development of deep learning-based models. The deep learning-based models show superior performance due to their ability to capture high-level semantic features automatically [7–11,31]. For instance, Ma et al. [7] proposed a recurrent neural network-based model to capture hidden temporal and textual features from relevant posts modeled as variable-length time series. Yu et al. [8] designed a convolutional neural network-based model to extract crucial textual features from the input sequence and model high-level interactions. Ma et al. [9] used the multi-task learning idea to jointly learn rumor detection and stance classification tasks to obtain better feature representation. Qi et al. [31] only utilized visual information from posts. They extracted visual features from the frequency and pixel domains and fused them to obtain feature representations of posts. In addition, researchers have also leveraged propagation structure features during the spread of rumors [12–16]. Concretely, Ma et al. [12] proposed a tree-structured recursive neural network to simultaneously capture textual semantic features and propagation structure features. Bian et al. [13] modeled the propagation and dispersion of rumors as top-down and bottom-up directed graphs, respectively, and leveraged graph convolutional networks to capture corresponding patterns and structure features. Sun et al. [16] modeled the dynamics of rumor propagation and background knowledge structures. Then, they utilized graph convolutional networks to capture these structure features at different time stages and incrementally combined them using a time fusion unit. However, such propagation structure features are generally unavailable at the early stage of rumor dissemination.

Although these unimodal methods have achieved good performance in rumor detection, they can only utilize features extracted from a single modality. The growing prevalence of multimodal posts on social platforms has amplified their limitations in effectively detecting multimodal rumors.

2.2. Multimodal Rumor Detection Methods

Existing multimodal rumor detection methods simultaneously utilize features extracted from both text and images to detect rumors. Some works added auxiliary tasks to the multimodal rumor detection task to improve detection performance [20–22]. Specifically, Wang et al. [20] designed an event adversarial neural network (EANN), which introduces an event discrimination task to capture event-invariant features by an event discriminator with the adversarial method. Khattar et al. [21] proposed a multimodal variational autoencoder (MVAE), which introduces a multimodal reconstruction loss to learn shared feature representation between textual and visual modalities by the variational autoencoder. Zhou et al. [22] devised a similarity-aware detection method called SAFE. They defined an extra detection loss based on the relevance score between modalities to help identify rumors, where the score is calculated through a modified cosine similarity algorithm. Moreover, with the tremendous success and advancement of pre-trained models, more and more methods extract multimodal features through pre-trained models for learning deeper and more complex semantics [23,24,26–30]. For instance, Singhal et al. [23,24] first used different pre-trained models to obtain both textual and visual feature representations. After that, using pre-trained models to extract multimodal features became the norm. In addition, considering that only employing a simple concatenation to fuse textual and visual features would discard associations between different modalities, many approaches explored more effective strategies to sufficiently utilize the relationships between modalities [25–30]. For example, Jin et al. [25] employed LSTM to obtain a joint representation of text and social context. Then, they designed a neuron-level attention mechanism to capture correlations between it and visual feature representations. Based on the better fusion of multimodal features by using the crossmodal weight-sharing layer and attention mechanism, Xue et al. [28] additionally considered visual tampering features and the semantic consistency of multimodal features. They defined an extra detection loss based on the cosine similarity between textual and visual semantic features to help detect rumors. Chen et al. [29] first defined an auxiliary correlation learning task to help achieve crossmodal feature alignment. Then, they adaptively aggregated unimodal and crossmodal fusion features based on an ambiguity score between modalities learned by estimating the KL divergence between distributions of textual and visual features. Moreover, Yang et al. [30] additionally considered multilevel modal features to obtain finer-grained representations. They used different encoding layers of BERT to capture different levels of textual semantic features and aggregated them in layers. Then, they used attention-based intramodal and intermodal fusion blocks to capture corresponding correlations, thus obtaining higher-order fusion features.

Although these multimodal rumor detection methods achieve relatively excellent performance, they mostly overlook the insufficient diversity of multimodal samples in feature space and hidden similarities and differences among multimodal samples, which limits the further improvement of their detection performance. Additionally, in recent years, contrastive learning has achieved remarkable performance improvements in various fields [32–37]. Meanwhile, the idea of data augmentation through interpolation has been proven to be effective [38]. With such considerations and inspired by these studies, we designed the Virtual Augmentation Contrastive Learning (VACL) module as an auxiliary training module to improve detection performance.

3. Methodology

3.1. Problem Definition

The multimodal rumor detection task is defined as a binary classification task, which aims to train a classifier to judge a given multimodal post as a rumor or a non-rumor. Formally, let $P = \{p_1, p_2, ..., p_n\}$ be a set of posts on social media for detection, where p_i is the i-th post and n is the number of posts. Each post $p_i = \{T_i, V_i, y_i\} \in P$ consists of the text content T_i, the attached image V_i, and the ground-truth label $y_i \in \{0, 1\}$ (i.e., non-rumor or rumor).

3.2. Overview

The overall architecture of the proposed MVACLNet is shown in Figure 1, which includes five modules: (**a**) a Hierarchical Textual Feature Extraction (HTFE) module, (**b**) a Visual Feature Extraction module, (**c**) a Multimodal Feature Fusion module, (**d**) a Virtual Augmentation Contrastive Learning (VACL) module, and (**e**) a Rumor Classification module. It should be noted that VACL, as an auxiliary training module for enhancing the feature learning of the model, only participates in model training. In the following sections, we will describe each component in detail. In particular, all feature representations in this paper are row vectors.

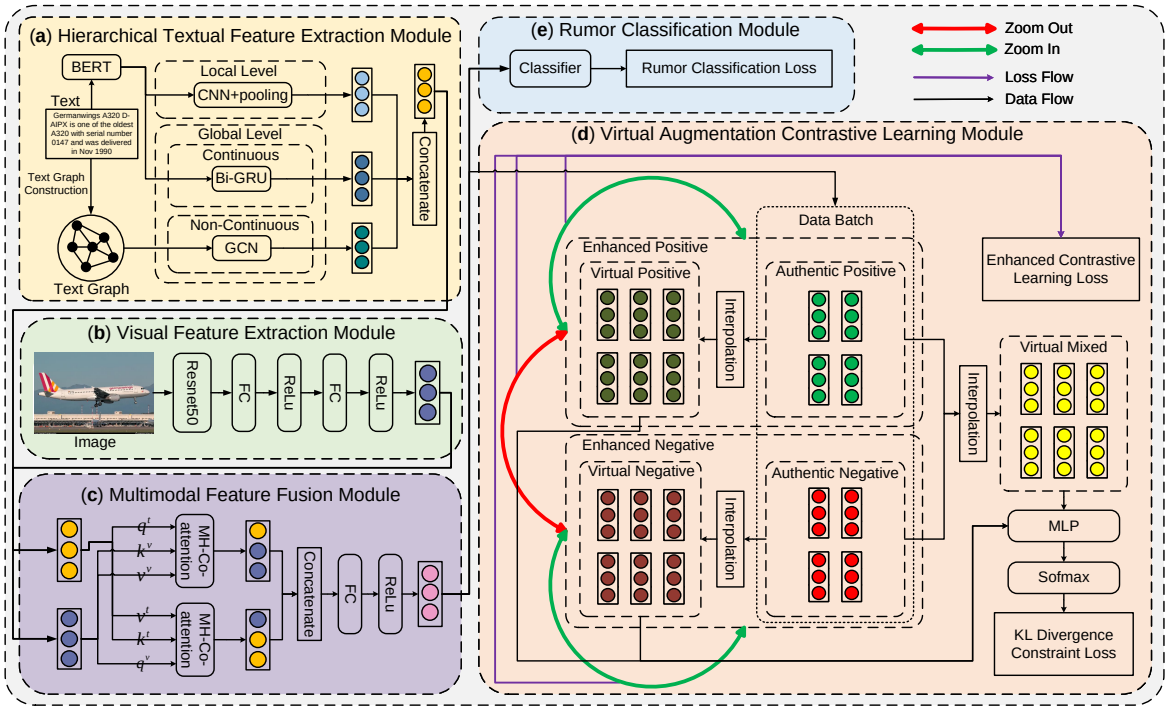

Figure 1. The overall architecture of MVACLNet.

3.3. Hierarchical Textual Feature Extraction Module

To comprehensively utilize crucial information in text data, we extract textual features from local and global levels, which contain three perspectives: local, global continuous, and global non-continuous. This approach contributes to the deep understanding of the meaning of the text and pays attention to its structure from different perspectives, overcoming the limitations of extracting a text's information from a single perspective. The process of HTFE is shown in Figure 1a.

Many studies have proven that the pre-trained language model BERT [39] has strong feature representation and transfer capabilities in various NLP tasks. To utilize its rich prior knowledge, we used BERT to initialize feature embeddings of each word w_i^t in the input text $T = \{w_1^t, w_2^t, ..., w_n^t\}$ to use in the local and global continuous perspectives:

$$H^t = \begin{bmatrix} h_1^t \\ h_2^t \\ ... \\ h_n^t \end{bmatrix} = \text{BERT}(T), \tag{1}$$

where $H^t \in \mathbb{R}^{n \times d^t}$ represents the word embedding matrix of T, h_i^t denotes the word embedding of w_i^t, n is the length of T, and d^t is the dimension of the word embeddings.

3.3.1. Local-Level Textual Feature Extraction

The local features of text reflect subtle semantic information and lexical associations well. We applied CNN with pooling [40] on H^t to capture such local features.

Specifically, a feature $m_{w,j}$ is first generated by applying a filter $W_w^{lt} \in \mathbb{R}^{d^t \times w}$ on a word embedding window $H_{i:i+w-1}^t$:

$$m_{w,j} = \text{ReLU}(H_{i:i+w-1}^t W_w^{lt} + b_w^{lt}), \tag{2}$$

where $j \in (1, n-w+1)$ represents the index of features obtained from the same convolutional kernel, w is the width of the convolutional kernel, and $b_w^{lt} \in \mathbb{R}$ is the bias term. This filter is applied to each possible word embedding window of H^t to generate a feature map:

$$m_w = [m_{w,1}, m_{w,2}, ..., m_{w,n-w+1}], \tag{3}$$

where $m_w \in \mathbb{R}^{n-w+1}$. Afterward, we apply the max-pooling operation on m_w to take the maximum value,

$$\hat{m}_w = \max(m_w) \tag{4}$$

as the feature corresponding to this filter, which aims to capture the most important feature for each feature map.

Based on the above steps of using a filter to extract a feature, multiple filters with different window sizes are employed to obtain multiple features. Here, we use d^{lt} filters with different receptive fields $w \in \{3, 4, 5\}$ to obtain the semantic features of different granularities, with an equal number of filters for each receptive field. All filters' outputs are concatenated to form the local-level textual feature representation $F^{lt} \in \mathbb{R}^{d^{lt}}$:

$$F^{lt} = \text{concat}(\hat{m}_{w=3,1}, ..., \hat{m}_{w=3,d^{lt}/3}, \hat{m}_{w=4,1}, ..., \hat{m}_{w=4,d^{lt}/3}, \hat{m}_{w=5,1}, ..., \hat{m}_{w=5,d^{lt}/3}), \tag{5}$$

where $\text{concat}(\cdot)$ represents a concatenation operation.

3.3.2. Global-Level Textual Feature Extraction

Textual Continuous Feature Extraction

The continuous features of text reflect contextual information and semantic relationships, which can help the model understand the meaning of text and its emotions. We apply BiGRU on H^t to extract such continuous features, which can further capture their temporal and sequential characteristics:

$$R^{gtc} = \begin{bmatrix} r_1 \\ r_2 \\ ... \\ r_n \end{bmatrix} = \text{BiGRU}(H^t), \tag{6}$$

where $R^{gtc} \in \mathbb{R}^{n \times d^{gtc}}$ represents the hidden state matrix of H^t; $r_i = [\overrightarrow{r_i}, \overleftarrow{r_i}] \in \mathbb{R}^{d^{gtc}}$ denotes the hidden state of h_i^t, which is formed by concatenating the forward hidden state $\overrightarrow{r_i} \in \mathbb{R}^{d^{gtc}/2}$ of h_i^t and the backward hidden state $\overleftarrow{r_i} \in \mathbb{R}^{d^{gtc}/2}$ of h_i^t.

Then, we concatenate $\overleftarrow{r_0}$ and $\overrightarrow{r_n}$ to obtain the global-level textual continuous feature representation $F^{gtc} \in \mathbb{R}^{d^{gtc}}$:

$$F^{gtc} = \text{concat}(\overleftarrow{r_0}, \overrightarrow{r_n}). \tag{7}$$

Textual Non-Continuous Feature Extraction

Social texts often exhibit characteristics of discretization and fragmentation. To better extract these non-continuous features from T, we transformed the feature extraction method into representation learning on a text graph.

Text Graph Construction.We converted each text into an independent heterogeneous text graph $G^t(V^t, E^t)$ with corresponding feature matrix X^t and adjacency matrix A^t, where

V^t represents the set of nodes that includes a text node and multiple word nodes, and E^t represents the edge set that includes text–word edges and word–word edges.

For constructing X^t, we used pre-trained feature embeddings to initialize feature representations of word nodes. The initial feature representation of the text node was obtained by accumulating feature embeddings of non-repeating words it contains. Formally, $X^t \in \mathbb{R}^{n^{tg} \times d^{gtnc}}$, where n^{tg} is the number of nodes in V^t, and d^{gtnc} denotes the dimension of the feature representation of nodes.

For constructing A^t, we built weights of text–word and word–word edges based on word occurrence in texts and word co-occurrence in the whole corpus [41]. Here, we employ the term frequency-inverse document frequency (TF-IDF) and positive point-wise mutual information (PPMI), respectively, to calculate the weights of the edges between the text nodes and the word nodes and between the word nodes. Formally, the weight of the edge between node n_i and node n_j is defined as

$$A^t_{i,j} = \begin{cases} \text{TF} - \text{IDF}(n_i, n_j) & n_i \text{ is text, } n_j \text{ is word} \\ \text{PPMI}(n_i, n_j) & n_i, n_j \text{ are words} \\ 1 & n_i = n_j \end{cases}. \tag{8}$$

The PPMI value between the word nodes is computed as

$$\text{PPMI}(n_i, n_j) = \max(\log \frac{p(n_i, n_j)}{p(n_i)p(n_j)}, 0), \tag{9}$$

$$p(n_i, n_j) = \frac{|W(n_i, n_j)|}{|W|}, \tag{10}$$

$$p(n_i) = \frac{|W(n_i)|}{|W|}, \tag{11}$$

where $|W|$ is the total number of sliding windows in the corpus, $|W(n_i)|$ is the number of sliding windows containing n_i, and $|W(n_i, n_j)|$ is the number of sliding windows containing both n_i and n_j. In particular, the statistical data used are based on the entire corpus rather than specific textual content.

Text Graph Representation Learning. After obtaining X^t and A^t, we employed a two-layer GCN [42] on the constructed text graph G^t to perform graph representation learning to obtain hidden representations of all nodes:

$$O^{gtnc} = \text{GCN}(X^t, A^t), \tag{12}$$

where $O^{gtnc} \in \mathbb{R}^{n^{tg} \times d^{gtnc}}$.

Finally, we applied global average pooling on O^{gtnc} to obtain the global-level textual non-continuous feature representation $F^{gtnc} \in \mathbb{R}^{d^{gtnc}}$. It is formulated as

$$F^{gtnc} = \text{MEAN}(O^{gtnc}). \tag{13}$$

3.3.3. Multi-Perspective Textual Feature Fusion

After obtaining F^{lt}, F^{gtc}, and F^{gtnc}, we concatenated them to obtain the final textual feature representation $F^t \in \mathbb{R}^{d^t}$:

$$F^t = \text{concat}(F^{lt}, F^{gtc}, F^{gtnc}), \tag{14}$$

where $d^t = d^{lt} + d^{gtc} + d^{gtnc}$.

3.4. Visual Feature Extraction Module

We used the pre-trained visual model ResNet50 [43] to extract the semantic features of the input image V. Specifically, we extracted the output of the penultimate layer of

ResNet50 and named it c^v, where $c^v \in \mathbb{R}^{d^{rno}}$ and d^{rno} is the output dimension of ResNet50. Then, we passed it through two fully connected layers with non-linear activation functions to obtain the visual feature representation F^v, that is,

$$F^v = \text{ReLu}(\text{ReLu}(c^v W_1^v + b_1^v)W_2^v + b_2^v), \tag{15}$$

where $F^v \in \mathbb{R}^{d^v}$ has the same dimension as the textual feature representation F^t, i.e., $d^v = d^t$, and W_*^v and b_*^v are corresponding linear transformation matrices and bias terms.

3.5. Multimodal Feature Fusion Module

Features of different modalities often carry distinct and complementary information. Effective integration contributes to improving a model's robustness and performance. To this end, we developed this module to effectively fuse the textual and visual features obtained from previous modules to form a richer and more precise multimodal feature representation. The process of the multimodal fusion scheme is shown in Figure 1c.

Specifically, we first performed multi-head linear mappings on the textual and visual feature representations F^t and F^v, respectively, to generate corresponding queries, keys, and values:

$$\begin{cases} q_h^t = F^t W_h^{q,t}, k_h^t = F^t W_h^{k,t}, v_h^t = F^t W_h^{v,t} \\ q_h^v = F^v W_h^{q,v}, k_h^v = F^v W_h^{k,v}, v_h^v = F^v W_h^{v,v} \end{cases}, \tag{16}$$

where all queries, keys, and values $\in \mathbb{R}^{d^h}$, all W_h^* are corresponding mapping matrices, h represents the index of attention heads, and d^h is the dimension of each attention head.

Then, we applied a co-attention mechanism [44], a variant of the self-attention mechanism [45], to produce enhanced crossmodal feature representations. Here, we perform it from two perspectives: textual-to-visual and visual-to-textual. Formally, the query comes from one modality, while the key and value are from another. In particular, considering that we need to calculate the attention weight matrix between a query and a key here, rather than between a set of queries and a set of keys like the original attention mechanism, we modified the calculation method of each attention score from a feature vector-based method to a feature value-based method for adapting to this situation, i.e., from the feature vector level to the feature value level, which could also help capture the correlation more finely. The computations are defined as

$$\begin{cases} z^{vt} = (\underset{h=1}{\overset{H}{\Big\|}} (v_h^t (\text{softmax}(\frac{(q_h^v)^T k_h^t}{\sqrt{d_h}}))^T))W^{o,vt} \\ z^{tv} = (\underset{h=1}{\overset{H}{\Big\|}} (v_h^v (\text{softmax}(\frac{(q_h^t)^T k_h^v}{\sqrt{d_h}}))^T))W^{o,tv} \end{cases}, \tag{17}$$

where $z^{vt} \in \mathbb{R}^{d^t}$ is the enhanced textual feature representation enhanced by the visual features; $z^{tv} \in \mathbb{R}^{d^v}$ is the enhanced visual feature representation enhanced by the textual features; $W^{o,vt} \in \mathbb{R}^{d^{mh} \times d^t}$; $W^{o,tv} \in \mathbb{R}^{d^{mh} \times d^v}$, d^{mh} is the feature dimension after concatenating features of multiple attention heads; and H is the number of attention heads.

Finally, we concatenated z^{vt} and z^{tv} together and passed them through a fully connected layer with a non-linear activation function to obtain the final multimodal feature representation $F \in \mathbb{R}^d$:

$$F = \text{ReLU}(\text{concat}(z^{vt}, z^{tv})W^{mff} + b^{mff}), \tag{18}$$

where $W^{mff} \in \mathbb{R}^{(d^t + d^v) \times d}$ denotes the linear transformation matrix, and $b^{mff} \in \mathbb{R}^d$ is the bias term.

3.6. Virtual Augmentation Contrastive Learning Module

Enhancing the diversity of multimodal samples in feature space and capturing hidden similarities and differences among multimodal samples can help the model adapt to a

broader range of complex data distributions and understand underlying data structures and patterns more effectively. With such considerations, we designed this module and used it as an auxiliary training module to assist the model in learning more robust and generalized multimodal feature representation to improve detection performance. The process of VACL is shown in Figure 1d.

To distinguish the virtual multimodal feature representations generated in this module, we use "authentic" to describe the multimodal feature representations that flow into VACL from the previous module, i.e., we rename the multimodal feature representation F as the authentic multimodal feature representation F^{auth}. Meanwhile, all labels are represented in a one-hot format.

3.6.1. Ground-Truth Label Introduction

We first introduced ground-truth labels to provide effective supervisory signals for subsequent contrastive learning, which can help prevent the model from misclassifying posts with noise and enhance contrastive learning.

In each data batch, let S^{auth} represent the authentic multimodal feature representation set, where $S^{auth} = \{F_i^{auth} | 1 \leq i \leq B\}$ and B is the size of each data batch. We first divided each authentic multimodal feature representation F_i^{auth} into authentic-positive multimodal feature representation $F_i^{auth,pos}$ or authentic-negative multimodal feature representation $F_i^{auth,neg}$ according to its ground-truth label. For example, when the ground-truth label of F_i^{auth} is "non-rumor", it is divided into $F_i^{auth,pos}$, and when the label of F_i^{auth} is "rumor", it is divided into $F_i^{auth,neg}$.

After that, we can obtain the authentic-positive multimodal feature representation set $S^{auth,pos}$ and the authentic-negative multimodal feature representation set $S^{auth,neg}$, where $S^{auth,pos} = \{F_i^{auth,pos} | 1 \leq i \leq N^{auth,pos}\}$, $S^{auth,neg} = \{F_i^{auth,neg} | 1 \leq i \leq N^{auth,neg}\}$, and $N^{auth,pos}$ and $N^{auth,neg}$ denote the number of authentic-positive and authentic-negative multimodal feature representations in set S^{auth}, respectively.

3.6.2. Virtual Sample Generation

Based on the above settings, to enhance the diversity of multimodal samples in feature space, we generated extra virtual multimodal feature representations and corresponding virtual labels using an interpolation-based data augmentation strategy [38], which were regarded as augmented samples for strengthening the feature learning of the model.

In this paper, we perform interpolation pairwise on the representations in set $S^{auth,pos}$, and all resulting representations form the virtual-positive multimodal feature representation set $S^{vir,pos}$. Similarly, we perform interpolation pairwise on the representations in set $S^{auth,neg}$, and all resulting representations compose the virtual-negative multimodal feature representation set $S^{vir,neg}$. Additionally, we perform interpolation pairwise, in a cross-set manner, on the representations in sets $S^{auth,pos}$ and $S^{auth,neg}$, and all resulting representations form the virtual-mixed multimodal feature representation set $S^{vir,mix}$. Specifically, the virtual-positive multimodal feature representation $F_k^{vir,pos} \in S^{vir,pos}$, the virtual-negative multimodal feature representation $F_k^{vir,neg} \in S^{vir,neg}$, and the virtual-mixed multimodal feature representation $F_k^{vir,mix} \in S^{vir,mix}$ are generated by the following formula:

$$\begin{cases} F_k^{vir,pos} = \lambda F_i^{auth,pos} + (1-\lambda) F_j^{auth,pos} \\ F_k^{vir,neg} = \lambda F_i^{auth,neg} + (1-\lambda) F_j^{auth,neg} \\ F_k^{vir,mix} = \lambda F_i^{auth,pos} + (1-\lambda) F_j^{auth,neg} \end{cases}, \tag{19}$$

where i, j, and k denote the indexes of these multimodal feature representations, i can be equal to j, and the interpolation balance parameter λ is sampled from the Beta distribution in each batch:

$$\lambda \sim \text{Beta}(\varepsilon, \varepsilon), \tag{20}$$

$$\lambda = \max(\lambda, 1-\lambda), \tag{21}$$

where ε is a hyperparameter used to control the distribution of λ.

Based on the same method, we generated corresponding virtual labels for each virtual multimodal feature representation, which formed three corresponding virtual label sets, $Y^{vir,pos}$, $Y^{vir,neg}$, and $Y^{vir,mix}$, where each virtual label $y_k^{vir,*}$ is generated as follows:

$$\begin{cases} y_k^{vir,pos} = \lambda y_i^{auth,pos} + (1-\lambda) y_j^{auth,pos} \\ y_k^{vir,neg} = \lambda y_i^{auth,neg} + (1-\lambda) y_j^{auth,neg} \\ y_k^{vir,mix} = \lambda y_i^{auth,pos} + (1-\lambda) y_j^{auth,neg} \end{cases}. \tag{22}$$

3.6.3. Sample Reorganization

To utilize these generated virtual multimodal feature representations and their corresponding virtual labels to perform subsequent feature learning, we reorganized them in advance.

Concretely, we merged $S^{auth,pos}$ and $S^{vir,pos}$ to compose the enhanced positive multimodal feature representation set $S^{enh,pos}$ and merged $S^{auth,neg}$ and $S^{vir,neg}$ to compose the enhanced negative multimodal feature representation set $S^{enh,neg}$.

Meanwhile, $S^{vir,pos}$, $S^{vir,neg}$, and $S^{vir,mix}$ were merged to form the virtual multimodal feature representation set S^{vir}. $Y^{vir,pos}$, $Y^{vir,neg}$, and $Y^{vir,mix}$ were merged to form the virtual label set Y^{vir}.

3.6.4. Enhanced Contrastive Learning

To catch more crucial similarities and differences among multimodal samples, we introduced contrastive learning and additionally leveraged ground-truth labels and the virtual multimodal feature representations to enhance it.

Specifically, for $S^{enh,pos}$ and $S^{enh,neg}$, we employed the contrastive learning strategy to encourage the multimodal feature representations belonging to the same class to be close to each other and those belonging to different classes to be far away from each other in the multimodal feature space. In this way, the model can capture more critical intra-class similarity features and inter-class difference features. The enhanced contrastive learning loss is defined as

$$\mathcal{L}^{con} = -\sum_{F_i^{con} \in S^{con}} \log \left\{ \frac{1}{|S^{pos}(F_i^{con})|} \sum_{F_j^{pos} \in S^{pos}(F_i^{con})} \frac{e^{\mathrm{sim}(F_i^{con}, F_j^{pos})/\tau}}{\sum_{F_k^{neg} \in S^{neg}(F_i^{con})} e^{\mathrm{sim}(F_i^{con}, F_k^{neg})/\tau}} \right\}, \tag{23}$$

where $S^{con} = S^{enh,pos} \cup S^{enh,neg}$; F_i^{con} represents the i-th multimodal feature representation in S^{con}; $S^{pos}(F_i^{con})$ denotes the positive sample set of F_i^{con}, which consists of those multimodal feature representations belonging to the same enhanced set as F_i^{con}; F_j^{pos} is the j-th positive sample of F_i^{con}; $S^{neg}(F_i^{con})$ represents the negative sample set of F_i^{con}, which is composed of those multimodal feature representations that do not belong to the same enhanced set as F_i^{con} ; F_k^{neg} is the k-th negative sample of F_i^{con}; the enhanced set is either $S^{enh,pos}$ or $S^{enh,neg}$; $|\cdot|$ denotes the counting operation on a set; $\mathrm{sim}(\cdot)$ represents the calculation function of cosine similarity; and $\tau \in \mathbb{R}^+$ is a scalar temperature parameter.

3.6.5. KL Divergence Constraint

For S^{vir} and Y^{vir}, we performed the KL divergence constraint, regarded as part of the total loss, between predicted probability distributions of the virtual multimodal feature representations and their corresponding virtual labels to help the model capture more content-invariant multimodal features:

$$\mathcal{L}^{kld} = \mathbb{E}_{F_i^{vir} \in S^{vir}, y_i^{vir} \in Y^{vir}} \mathrm{KL}(\mathrm{Softmax}(\mathrm{MLP}(F_i^{vir})) || y_i^{vir}), \tag{24}$$

where F_i^{vir} is the i-th virtual multimodal feature representation in S^{vir}, y_i^{vir} denotes the virtual label of F_i^{vir}, and a two-layer MLP and a softmax function are employed to predict the probability distribution of F_i^{vir}.

3.7. Rumor Classification Module

We performed rumor classification on the authentic multimodal feature representations in S^{auth}. The predicted label $\hat{y}_i$ of $F_i^{auth} \in S^{auth}$ was calculated via a fully connected layer and a softmax function:

$$\hat{y}_i = \text{softmax}(F_i^{auth} W + b), \tag{25}$$

where W and b are the weight and bias parameters. Then, we used the cross-entropy loss as the rumor classification loss:

$$\mathcal{L}^{cls} = -\sum_i y_i \log(\hat{y}_i), \tag{26}$$

where y_i denotes the ground-truth label of F_i^{auth}.

3.8. Overall Loss

The overall loss is a weighted sum of the rumor classification loss, the enhanced contrastive learning loss, and the KL divergence constraint loss:

$$\mathcal{L} = \alpha \mathcal{L}^{cls} + \beta \mathcal{L}^{con} + \gamma \mathcal{L}^{kld}, \tag{27}$$

where α, β, and γ are hyperparameters that balance these three losses.

We optimized the parameters of the model by minimizing the overall loss $\mathcal{L}$.

4. Experiments and Analysis

In this section, we first introduce the datasets used, the experimental setup, and the baseline models. Second, we compare the detection performance of MVACLNet with the baseline models. Then, we perform the ablation analysis to verify the effectiveness of each module and component of MVACLNet. Finally, we visually analyze the role of VACL.

4.1. Datasets

We conducted experiments on two real-world datasets, Twitter [46] and Weibo [25], to verify the effectiveness of the proposed MVACLNet.

In the Twitter dataset, each tweet contains textual content, image/video, and associated social context information. The dataset has around 17,000 unique tweets spanning different events. The dataset was divided into two parts: the development set (9000 rumor tweets, 6000 non-rumor tweets) and the test set (2000 tweets). There are no overlapping events between these two sets. In particular, all the tweets were manually verified by cross-checking online sources (articles and blogs). Due to our focus on using text and image content to identify rumors, we removed posts that only attached videos. We used the development set for training and the test set for testing.

In the Weibo dataset, each post contains textual content, image content, and social context information. The dataset contains 4749 rumors and 4779 non-rumors. The non-rumor posts were all verified by the Xinhua News Agency, an authoritative news agency in China, and rumors were all confirmed by the official rumor debunking system of Weibo. We only focus on textual and image information, ignoring social context information. The dataset is divided into the training and testing sets in an 8:2 ratio.

4.2. Experimental Setup

For the model implementation and training, our algorithms were implemented using the Pytorch framework [47] and trained with Adam Optimizer [48].

For the data preprocessing, we removed URLs and emoticons from text and uniformly set the size of images to 224 × 224 × 3.

For the settings of the pre-trained models, we used bert-base-uncased for English text and bert-base-chinese for Chinese text. In particular, we kept the parameters of BERT and ResNet50 static to avoid overfitting.

For the initial word embeddings in the global-level textual non-continuous feature extraction component, we used GloVe [49] for English and FastText [50] for Chinese.

For the dimension settings, the word embedding dimension initialized by BERT was 768; the output dimension of the penultimate layer of ResNet50 was 2048; the dimensions d^{lt}, d^{gtc}, and d^{gtnc} were all 300; the dimensions d^{t} and d^{v} were both 900; and the dimension d was 300.

For the parameter settings, we set the learning rate to 0.002, the batch size to 64, and the number of attention heads, i.e., H, to 9. The hyperparameter ε, which controls the distribution of λ, was set to 0.7 for both Twitter and Weibo. The hyperparameters α, β, and γ, which balance losses, were set to 1.0, 1.0, and 1.5 for Twitter, respectively, and 1.0, 2.0, and 1.5 for Weibo, respectively.

For the training strategies, we adopted the early stop strategy and the dynamic learning rate reduction strategy during the model training and set the value of dropout used for attention to 0.1 to avoid overfitting.

For the evaluation metrics, we used accuracy, precision, recall, and F1 score as the evaluation metrics to assess experimental performance, as these metrics are widely accepted for evaluating multimodal rumor detection methods. All the aforementioned evaluation metrics are represented as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{28}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{29}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{30}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{31}$$

where TP (true positive) represents the number of positive samples classified correctly as positive, TN (true negative) represents the number of negative samples classified correctly as negative, FP (false positive) represents the number of negative samples incorrectly classified as positive, and FN (false negative) represents the number of positive samples incorrectly classified as negative.

4.3. Baselines

To evaluate the effectiveness and superiority of MVACLNet, we compare it with two types of baseline models: unimodal models and multimodal models, where the former is based on textual or visual content in the post and the latter relies on both textual and visual content in the post.

(a) Unimodal Models:

- VGG-19 [51]: This model is a pre-trained deep convolutional neural network architecture with 19 layers, which is widely employed for image classification tasks and is known for its straightforward yet effective stacked convolutional layer structure. We used it to obtain visual feature representation, which is input into a fully connected layer followed by a softmax layer to detect rumors.
- BERT [39]: This model is a pre-trained language model based on the Transformer encoder architecture, which captures bidirectional contextual textual information. We used it to obtain textual feature representation, which is input into a fully connected layer followed by a softmax layer to detect rumors.

(b) Multimodal Models:

- att-RNN [25]: This model employs LSTM to learn a joint representation of text and social context, and it extracts visual features through VGG-19. Then, it designs a neuron-level attention mechanism to capture correlations between visual and textual social features to obtain an attention-aggregated visual representation. Finally, it concatenates the representations for rumor detection. For fairness in the comparison, we removed the part that deals with social features in the concrete implementation.
- EANN [20]: This model uses TextCNN [40] and VGG-19 to extract textual and visual features, respectively. Then, it concatenates them as a multimodal feature representation, which is input into an event discriminator and a rumor classifier. The event discriminator guides the model to capture event-invariant multimodal features through an event adversarial mechanism.
- MVAE [21]: This model utilizes the variational autoencoder and a designed multimodal reconstruction loss to learn a shared representation between textual and visual modalities, where the encoder extracts textual and visual features through bidirectional LSTM and VGG19, respectively. Finally, the sampled latent multimodal feature representation is used for rumor detection.
- Spotfake [23]: This model uses BERT to extract textual features and utilizes VGG19 to capture visual features. Then, it concatenates them for rumor detection.
- Spotfake+ [24]: This model is an upgraded version of SpotFake that replaces BERT with the pre-trained language model XLNet [52] to extract textual features.
- SAFE [22]: This model first converts an image to text through a pre-trained image2sentence model. Then, it uses TextCNN to capture textual and visual features, which are concatenated for rumor detection. Meanwhile, it further utilizes the relevance between modalities, quantified as crossmodal similarity, to define an extra detection loss, thereby helping identify rumors.
- MCNN [28]: This model applies BERT and BiGRU to capture textual semantic features and utilizes ResNet50, an attention mechanism, and BiGRU to extract visual semantic features. Meanwhile, it captures visual tampering features through the Error Level Analysis (ELA) algorithm and ResNet50. Then, it uses a crossmodal weight-sharing layer and the attention mechanism to fuse all the above features and the semantic features of ResNet50 output for prediction. Afterward, it further employs cosine similarity to measure the similarity between textual and visual semantic features, which is used to define an extra detection loss to help detect rumors.
- CAFE [29]: This model utilizes BERT and ResNet34 to extract textual and visual features, respectively. Then, it defines an auxiliary correlation learning task to help achieve crossmodal feature alignment. Following this, it adaptively aggregates unimodal features and crossmodal correlations based on a learned ambiguity score between modalities, where the score is quantified by estimating the KL divergence between distributions of textual and visual features. Finally, the aggregated multimodal feature representation is input into a classifier for rumor detection.
- MRAN [30]: This model first extracts multilevel textual semantic features through different encoding layers of BERT. Then, it further utilizes TextCNN to aggregate these features in layers, thus filtering out some noise while extracting important local information. The visual features are extracted through VGG-19. Afterward, it uses text/image attention blocks and cross-attention blocks to capture intramodal and intermodal associations, thereby obtaining higher-order fusion features between textual and visual modalities. Finally, the fused multimodal feature representation is used for rumor detection.

4.4. Comparative Experiments and Analysis

Tables 1 and 2 exhibit the detection performance of MVACLNet and all baseline models on the two real-world datasets, where the experimental results of the baseline models were

obtained on the same datasets according to their model structure and experimental setup. It should be noted that all experimental results are from the test set.

Table 1. The performance of different methods on the Twitter dataset.

Method	Accuracy	Rumor			Non-Rumor		
		Precision	Recall	F1	Precision	Recall	F1
VGG-19	0.596	0.695	0.518	0.593	0.524	0.700	0.599
BERT	0.706	0.648	0.540	0.589	0.715	0.636	0.673
att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
Spotfake	0.771	0.784	0.744	0.764	0.769	0.807	0.787
Spotfake+	0.790	0.793	0.827	0.810	0.786	0.747	0.766
MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
CAFE	0.806	0.807	0.799	0.803	0.805	0.813	0.809
MRAN	0.855	**0.861**	0.857	0.859	0.847	0.816	0.831
MVACLNet	**0.891**	0.811	**0.922**	**0.863**	**0.949**	**0.872**	**0.909**

The bold numbers indicate the optimal values for the corresponding metrics.

Table 2. The performance of different methods on the Weibo dataset.

Method	Accuracy	Rumor			Non-Rumor		
		Precision	Recall	F1	Precision	Recall	F1
VGG-19	0.633	0.630	0.500	0.550	0.630	0.750	0.690
BERT	0.804	0.800	0.860	0.830	0.840	0.760	0.800
att-RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
Spotfake	0.869	0.877	0.859	0.868	0.861	0.879	0.870
Spotfake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873
MCNN	0.846	0.809	0.857	0.832	0.879	0.837	0.858
CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
MRAN	0.903	0.904	0.908	0.906	0.897	0.892	0.894
MVACLNet	**0.913**	**0.916**	**0.911**	**0.913**	**0.910**	**0.916**	**0.913**

The bold numbers indicate the optimal values for the corresponding metrics.

From the experimental results of unimodal models, we can observe that BERT outperforms VGG-19, which shows that textual features are more important and influential than visual features for the rumor detection task.

For the comparison between unimodal and multimodal models, BERT performs better than att-RNN and EANN on the Twitter dataset, and it also obtains better performance than att-RNN, EANN, and SAFE on the Weibo dataset, which indicates the advantage of utilizing BERT to extract textual semantic features to detect rumors. However, on the whole, its performance is still lower than that of other multimodal models, which shows the limitation of using unimodal models to detect multimodal rumors. These results also confirm the effectiveness of visual content as supplementary information and the importance of designing an effective Multimodal Feature Fusion schema for multimodal rumor detection.

From the experimental results of multimodal models, we can see that Spotfake and Spotfake+ outperform att-RNN, EANN, MVAE, and SAFE on the two datasets, which proves the advantage of using different pre-trained models to extract multimodal features in the multimodal rumor detection task. MCNN and CAFE achieve better performance than att-RNN, EANN, MVAE, and SAFE on the two datasets. At the same time, CAFE

performs better than Spotfake and Spotfake+, and MCNN performs more satisfactorily than Spotfake on the Twitter dataset. Moreover, MRAN obtains the best detection performance of all the multimodal baseline models on the two datasets. These findings illustrate that capturing finer-grained and comprehensive modal features, leveraging intramodal and intermodal associations, and designing more effective feature fusion strategies between modalities can benefit the multimodal rumor detection task.

For our proposed model, the experimental results show that MVACLNet outperforms all the baseline models on these two datasets in terms of accuracy and most other metrics, confirming its effectiveness and superiority. These results also demonstrate that additionally considering extracting textual features from multiple perspectives to comprehensively utilize textual information, enhancing the diversity of multimodal samples in feature space to enhance feature learning, capturing more crucial similarities and differences among multimodal samples, and extracting more content-invariant multimodal features can be further beneficial to the multimodal rumor detection task, thus improving detection performance.

Furthermore, we can observe that these models generally perform better on the Weibo dataset than on the Twitter dataset. This discrepancy may arise due to several distinctions between these two datasets. In terms of textual content, the average post length in the Twitter dataset is smaller than that in the Weibo dataset, making it unable to provide more textual semantic information. Regarding image content, the number of images in the Twitter dataset is less than the number of posts, indicating that there are instances where a single image is shared by multiple posts, resulting in insufficient image features in the Twitter dataset. In contrast, each post in the Weibo dataset has one or multiple corresponding images. These dissimilarities inevitably lead to variations in the performance of each model when applied to these two datasets.

4.5. Ablation Experiments and Analysis

To assess the effectiveness of each module or component in MVACLNet, we designed several variants to investigate the impacts of these modules or components on MVACLNet:

- "w/o VACL" represents a model without the Virtual Augmentation Contrastive Learning module.
- "w/o KLC" represents a model that only uses the enhanced contrastive learning without performing the KL divergence constraint in VACL.
- "w/o ECL" indicates a model that only implements the KL divergence constraint without employing the enhanced contrastive learning in VACL.
- "w/o KLC+VA" represents a model that only leverages ground-truth labels to enhance contrastive learning in VACL, which does not perform the KL divergence constraint and does not use the extra-generated virtual multimodal feature representations to enhance contrastive learning.
- "w/o VA" represents a model that only leverages ground-truth labels to enhance contrastive learning and performs the KL divergence constraint in VACL, which does not use the extra-generated virtual multimodal feature representations to enhance contrastive learning.
- "w/o MFF" signifies a model with the Multimodal Feature Fusion module replaced by a two-layer MLP with an ReLu activation function.
- "w/o LT" represents a model that removes the local-level textual feature extraction component from HTFE.
- "w/o GTC" signifies a model that removes the global-level textual continuous feature extraction component from HTFE.
- "w/o GTNC" denotes a model that removes the global-level textual non-continuous feature extraction component from HTFE.

Tables 3 and 4 present the performance of variants of MVACLNet on the two real-world datasets. It should be noted that all experimental results are from the test set.

Table 3. The performance of different variants of MVACLNet on the Twitter dataset.

Method	Accuracy	Rumor			Non-Rumor		
		Precision	Recall	F1	Precision	Recall	F1
MVACLNet	**0.891**	**0.811**	**0.922**	**0.863**	**0.949**	0.872	**0.909**
w/o VACL	0.824	0.763	0.763	0.763	0.860	0.860	0.860
w/o ECL	0.831	0.772	0.776	0.774	0.867	0.864	0.865
w/o KLC	0.843	0.792	0.781	0.787	0.872	0.879	0.875
w/o KLC+VA	0.847	0.770	0.837	0.802	0.899	0.853	0.875
w/o VA	0.857	0.768	0.883	0.821	0.923	0.841	0.881
w/o MFF	0.857	0.772	0.871	0.818	0.918	0.849	0.882
w/o LT	0.833	0.746	0.831	0.786	0.894	0.834	0.863
w/o GTC	0.841	0.797	0.767	0.781	0.866	**0.885**	0.875
w/o GTNC	0.830	0.763	0.784	0.773	0.870	0.857	0.864

The bold numbers indicate the optimal values for the corresponding metrics.

Table 4. The performance of different variants of MVACLNet on the Weibo dataset.

Method	Accuracy	Rumor			Non-Rumor		
		Precision	Recall	F1	Precision	Recall	F1
MVACLNet	**0.913**	**0.916**	**0.911**	**0.913**	**0.910**	**0.916**	**0.913**
w/o VACL	0.874	0.876	0.872	0.874	0.872	0.876	0.874
w/o ECL	0.879	0.890	0.865	0.877	0.868	0.892	0.880
w/o KLC	0.869	0.880	0.857	0.868	0.860	0.882	0.871
w/o KLC+VA	0.876	0.882	0.869	0.876	0.870	0.883	0.876
w/o VA	0.878	0.890	0.862	0.876	0.866	0.893	0.879
w/o MFF	0.874	0.900	0.842	0.870	0.851	0.906	0.878
w/o LT	0.874	0.853	0.907	0.879	0.900	0.842	0.870
w/o GTC	0.879	0.893	0.861	0.877	0.865	0.896	0.880
w/o GTNC	0.883	0.868	0.903	0.885	0.899	0.862	0.880

The bold numbers indicate the optimal values for the corresponding metrics.

Based on the results of the ablation experiments, it is evident that each module or component serves a distinct purpose, and removing any module or component will impact the overall performance of MVACLNet.

For "w/o VACL", compared with the complete model, its accuracy decreased by 6.7% and 3.9% on the Twitter and Weibo datasets, respectively. Meanwhile, its F1 scores for the rumor and non-rumor categories decreased by 10% and 4.9%, respectively, on the Twitter dataset, and both decreased by 3.9% on the Weibo dataset. These results indicate that removing the VACL module has a noticeable effect on the model's performance, affirming the effectiveness and generalization of the VACL module proposed in this paper.

From the experimental results of "w/o ECL", "w/o w/o KLC", "w/o KLC+VA", and "w/o VA", we can observe that the accuracy and the F1 scores of these variants all decreased to varying degrees compared with the complete model. These findings suggest that the lack of any component of VACL does not yield optimal performance. Instead, the joint consideration of enhancing the diversity of multimodal samples in feature space to strengthen feature learning, capturing more crucial similarities and differences among multimodal samples, and extracting more content-invariant multimodal features can complement and enhance each other to help the model learn more robust and generalized multimodal feature representations, thereby achieving superior detection performance. These results also indirectly prove that the extra-generated virtual multimodal feature representations and the introduced ground-truth labels can enhance contrastive learning to capture more critical intra-class similarity features and inter-class difference features.

For "w/o MFF", its accuracy decreased by 3.4% and 3.9% on the Twitter and Weibo datasets, respectively, compared with the complete model. Meanwhile, its F1 scores for the rumor and non-rumor categories decreased by 4.5% and 2.7%, respectively, on the Twitter dataset and decreased by 4.3% and 3.5%, respectively, on the Weibo dataset.

These results show that the proposed Multimodal Feature Fusion module effectively fuses multimodal features and explores the underlying connections between textual and visual features, which also proves that designing an effective feature fusion strategy contributes to multimodal rumor detection.

The experimental results of "w/o LT", "w/o GTC", and "w/o GTNC" all exhibited varying degrees of decreases compared to the complete model, which illustrates that neglecting textual features from any perspective does not achieve optimal detection performance. These findings demonstrate the effectiveness of our proposed HTFE module, i.e., considering extracting textual features from multiple perspectives to comprehensively utilize textual information can benefit multimodal rumor detection.

4.6. Visualization Analysis

To further explore and analyze the effectiveness of the proposed VACL module, we used the t-SNE algorithm [53] to visually compare the multimodal feature representations learned by "w/o VACL" with those learned by MVACLNet on the two datasets. The t-SNE algorithm maps high-dimensional data to a two-dimensional space, which are presented on a two-dimensional coordinate map for visualization. The visualization results are shown in Figure 2.

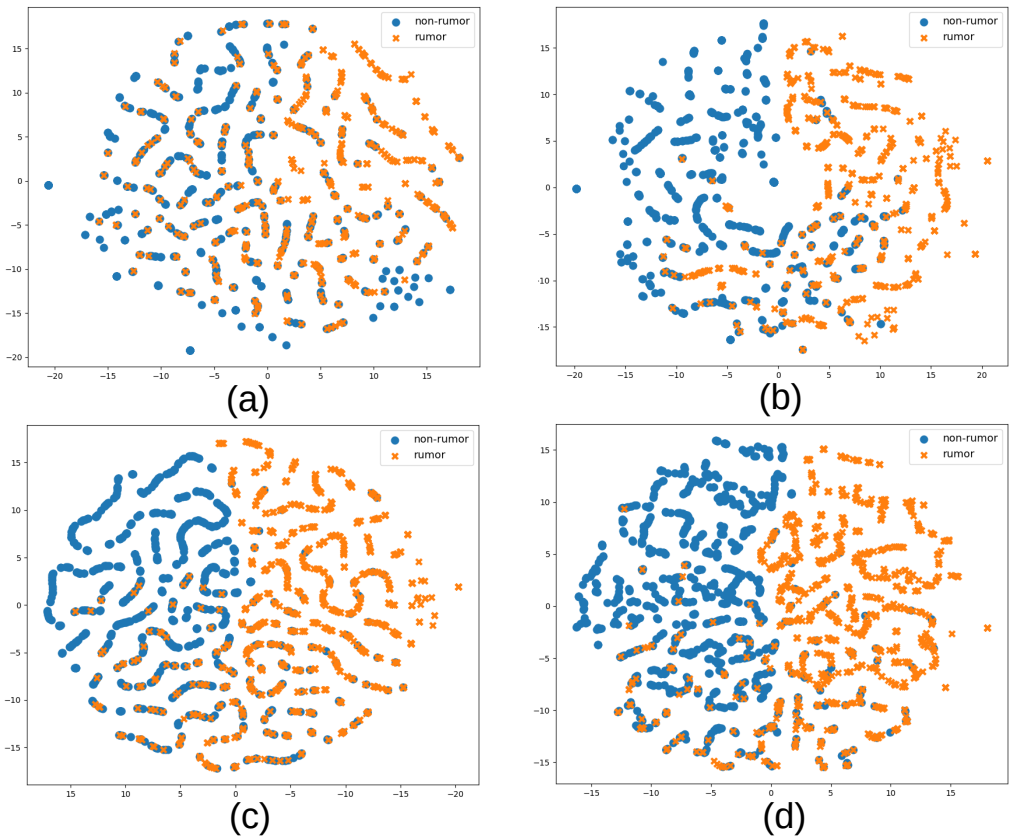

Figure 2. The visualization results of the multimodal feature representations learned by "w/o VACL" and MVACLNet on the Twitter and Weibo datasets. (**a**) The visualization result of the multimodal feature representations learned by "w/o VACL" on the Twitter dataset. (**b**) The visualization result of the multimodal feature representations learned by MVACLNet on the Twitter dataset. (**c**) The visualization result of the multimodal feature representations learned by "w/o VACL" on the Weibo dataset. (**d**) The visualization result of the multimodal feature representations learned by MVACLNet on the Weibo dataset.

In Figure 2, we can observe that Figure 2a,c contain more misclassified samples compared to Figure 2b,d. In addition, the interval of the distribution for sample points in

Figure 2b,d are more distinct than those in Figure 2a,c, with clustering results being more concentrated. These results show that MVACLNet exhibits more satisfactory classification performance, which demonstrates that MVACLNet outperforms "w/o VACL" in learning multimodal feature representations and further validates the effectiveness of the VACL module. The reason is that, compared with "w/o VACL", MVACLNet can enhance feature learning by enhancing the diversity of multimodal samples in feature space and capture more critical similarities and differences among multimodal samples, more content-invariant multimodal features, and more representative latent multimodal features of rumors. These help the model acquire more robust and generalized multimodal feature representations, thus improving the performance of multimodal rumor detection.

5. Limitations and Threats to Validity

Our model performs better than the baseline models on the Twitter and Weibo datasets, while its performance on other datasets still needs further validation. Our model's performance on strong noise datasets also needs further testing.

6. Conclusions

In this paper, we propose a Multimodal Virtual Augmentation Contrastive Learning Network (MVACLNet) for rumor detection. First, MVACLNet utilizes the designed Hierarchical Textual Feature Extraction (HTFE) module to extract textual features from multiple perspectives, thereby leveraging textual information comprehensively. Second, it better integrates the textual and virtual features using a modified cross-attention mechanism, which operates from different perspectives at the feature value level, to obtain richer and more precise multimodal feature representations. Third, a Virtual Augmentation Contrastive Learning (VACL) module is devised as an auxiliary training module to help the model learn more robust and generalized multimodal feature representations by enhancing the diversity of multimodal samples in feature space for enhancing feature learning, capturing more critical similarities and differences among multimodal samples, and extracting more content-invariant multimodal features. The experimental results demonstrate the effectiveness and superiority of MVACLNet on multimodal rumor detection.

In future work, we aim to explore an improved multimodal learning method suitable for detecting multimodal rumors with a certain extent of propagation structure, as in some cases, multimodal rumors may have already spread over a period of time and formed a certain scale of information dissemination structure.

Author Contributions: Conceptualization, X.L. and M.P.; methodology, X.L. and M.P.; validation, X.L. and M.P.; writing—original draft preparation, X.L. and M.P.; writing—review and editing, X.L., M.P., Q.L., J.Z., H.W. and D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Shandong Province, grant number ZR2020MF045, and the Key Research and Development (R&D) Plan of Shandong Province (Soft Science), grant number 2023RKL01004.

Data Availability Statement: We used publicly available datasets for all the experiments carried out in this paper. The Twitter dataset is available at GitHub (accessed on 4 May 2023): https://github.com/MKLab-ITI/image-verification-corpus. The Weibo dataset is available at GitHub (accessed on 4 May 2023): https://github.com/wangzhuang1911/Weibo-dataset.

Acknowledgments: The authors are very thankful to the editor and referees for their valuable comments and suggestions for improving the paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

MVACLNet	Multimodal Virtual Augmentation Contrastive Learning Network.
HTFE	Hierarchical Textual Feature Extraction.
VACL	Virtual Augmentation Contrastive Learning.
KL	Kullback–Leibler.
LSTM	Long Short-Term Memory.
BERT	Bidirectional Encoder Representation from Transformers.
CNN	convolutional neural network.
BiGRU	Bidirectional Gated Recurrent Unit.
TF-IDF	Term Frequency-Inverse Document Frequency.
PPMI	Positive Point-Wise Mutual Information.
GCN	graph convolutional network.
ResNet	Residual Network.
VGG	Visual Geometry Group.
att-RNN	Recurrent Neural Network with an attention mechanism.
EANN	Event Adversarial Neural Network.
MVAE	multimodal variational autoencoder.
SAFE	Similarity-Aware FakE news detection method.
MCNN	Multimodal Consistency Neural Network.
MRAN	Multimodal Relationship-Aware Attention Network.
TP	true positive.
TN	true negative.
FP	false positive.
FN	false negative.

References

1. Zhang, X.; Ghorbani, A.A. An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manag.* **2020**, *57*, 102025. [CrossRef]
2. Naeem, S.B.; Bhatti, R.; Khan, A. An exploration of how fake news is taking over social media and putting public health at risk. *Health Inf. Libr. J.* **2021**, *38*, 143–149. [CrossRef] [PubMed]
3. Castillo, C.; Mendoza, M.; Poblete, B. Information credibility on twitter. In Proceedings of the 20th International World Wide Web Conference, Hyderabad, India, 28 March–1 April 2011; pp. 675–684.
4. Zhao, Z.; Resnick, P.; Mei, Q. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. In Proceedings of the 24th International World Wide Web Conference, Florence, Italy, 18–22 May 2015; pp. 1395–1405.
5. Jin, Z.; Cao, J.; Zhang, Y.; Luo, J. News verification by exploiting conflicting social viewpoints in microblogs. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2972–2978.
6. Shao, C.; Ciampaglia, G.L.; Flammini, A.; Menczer, F. Hoaxy: A platform for tracking online misinformation. In Proceedings of the 25th International World Wide Web Conference, Montréal, QC, Canada, 11–15 April 2016; pp. 745–750.
7. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B.J.; Wong, K.F.; Cha, M. Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 3818–3824.
8. Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T. A Convolutional Approach for Misinformation Identification. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 3901–3907.
9. Ma, J.; Gao, W.; Wong, K.F. Detect rumor and stance jointly by neural multi-task learning. In Proceedings of the Companion Proceedings of the Web Conference 2018, Lyon, France, 23–27 April 2018; pp. 585–593.
10. Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; Li, J. MDFEND: Multi-domain fake news detection. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual Event, 1–5 November 2021; pp. 3343–3347.
11. Wu, L.; Rao, Y.; Zhang, C.; Zhao, Y.; Nazir, A. Category-controlled encoder-decoder for fake news detection. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 1242–1257. [CrossRef]
12. Ma, J.; Gao, W.; Wong, K.F. Rumor detection on Twitter with tree-structured recursive neural networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1980–1989.
13. Bian, T.; Xiao, X.; Xu, T.; Zhao, P.; Huang, W.; Rong, Y.; Huang, J. Rumor detection on social media with bi-directional graph convolutional networks. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 549–556.
14. Wei, L.; Hu, D.; Zhou, W.; Yue, Z.; Hu, S. Towards propagation uncertainty: Edge-enhanced bayesian graph convolutional networks for rumor detection. *arXiv* **2021**, arXiv:2107.11934.

15. Hu, D.; Wei, L.; Zhou, W.; Huai, X.; Han, J.; Hu, S. A rumor detection approach based on multi-relational propagation tree. *J. Comput. Res. Dev.* **2021**, *58*, 1395–1411.
16. Sun, M.; Zhang, X.; Zheng, J.; Ma, G. DDGCN: Dual Dynamic Graph Convolutional Networks for Rumor Detection on Social Media. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, Virtual Event, 22 February–1 March 2022; pp. 4611–4619.
17. Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; Tian, Q. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Trans. Multimed.* **2017**, *19*, 598–608. [CrossRef]
18. Alam, F.; Cresci, S.; Chakraborty, T.; Silvestri, F.; Dimitrov, D.; Martino, G.D.S.; Shaar, S.; Firooz, H.; Nakov, P. A survey on multimodal disinformation detection. *arXiv* **2021**, arXiv:2103.12541.
19. Silva, A.; Luo, L.; Karunasekera, S.; Leckie, C. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 557–565.
20. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 849–857.
21. Khattar, D.; Goud, J.S.; Gupta, M.; Varma, V. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2915–2921.
22. Zhou, X.; Wu, J.; Zafarani, R. SAFE: Similarity-Aware Multi-modal Fake News Detection. In Proceedings of the 24th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Singapore, 11–14 May 2020; pp. 354–367.
23. Singhal, S.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S.I. Spotfake: A multi-modal framework for fake news detection. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; pp. 39–47.
24. Singhal, S.; Kabra, A.; Sharma, M.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 13915–13916.
25. Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; Luo, J. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 795–816.
26. Qian, S.; Wang, J.; Hu, J.; Fang, Q.; Xu, C. Hierarchical multi-modal contextual attention network for fake news detection. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 11–15 July 2021; pp. 153–162.
27. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. Multimodal fusion with co-attention networks for fake news detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Bangkok, Thailand, 1–6 August 2021; pp. 2560–2569.
28. Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; Wei, L. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manag.* **2021**, *58*, 102610. [CrossRef] [PubMed]
29. Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; Shang, L. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In Proceedings of the ACM Web Conference 2022, Virtual Event, 25–29 April 2022; pp. 2897–2905.
30. Yang, H.; Zhang, J.; Zhang, L.; Cheng, X.; Hu, Z. MRAN: Multimodal relationship-aware attention network for fake news detection. *Comput. Stand. Interfaces* **2024**, *89*, 103822. [CrossRef]
31. Qi, P.; Cao, J.; Yang, T.; Guo, J.; Li, J. Exploiting Multi-domain Visual Information for Fake News Detection. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; pp. 518–527.
32. Dai, B.; Lin, D. Contrastive Learning for Image Captioning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 898–907
33. Cai, H.; Chen, H.; Song, Y.; Ding, Z.; Bao, Y.; Yan, W.; Zhao, X. Group-wise Contrastive Learning for Neural Dialogue Generation. *arXiv* **2020**, arXiv:2009.07543.
34. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 1597–1607.
35. Wu, H.; Ma, T.; Wu, L.; Manyumwa, T.; Ji, S. Unsupervised Reference-Free Summary Quality Evaluation via Contrastive Learning. *arXiv* **2020**, arXiv:2010.01781.
36. Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; Wang, L. Graph Contrastive Learning with Adaptive Augmentation. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 2069–2080.
37. Sun, T.; Qian, Z.; Dong, S.; Li, P.; Zhu, Q. Rumor Detection on Social Media with Graph Adversarial Contrastive Learning. In Proceedings of the ACM Web Conference 2022, Virtual Event, 25–29 April 2022; pp. 2789–2797.
38. Chen, J.; Yang, Z.; Yang, D. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. *arXiv* **2020**, arXiv:2004.12239.
39. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
40. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:1408.5882.

41. Yao, L.; Mao, C.; Luo, Y. Graph Convolutional Networks for Text Classification. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 7370–7377.
42. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
44. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 12–23.
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
46. Boididou, C.; Papadopoulos, S.; Zampoglou, M.; Apostolidis, L.; Papadopoulou, O.; Kompatsiaris, Y. Detection and visualization of misleading content on Twitter. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 71–86. [CrossRef]
47. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
48. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
49. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
50. Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning word vectors for 157 languages. *arXiv* **2018**, arXiv:1802.06893.
51. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
52. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5754–5764.
53. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

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.