THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
電機及電子工程學系

PhD Thesis Defense

# Deep Learning for Rumour Detection and Claim Veracity Assessment on Social Media

**CHEUNG Tsun Hin**

Supervisor: Prof. LAM Kin Man

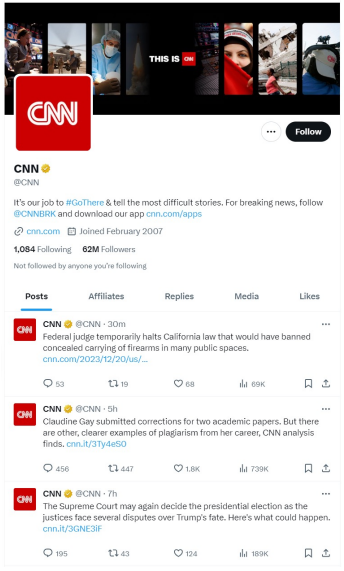
Date: 28 December 2023

# Introduction and Background

# Online Social Media

**Social media** has become a convenient way for people to exchange information.

- Twitter (rebranded to X after Elon Musk's acquisition in 2023.)
- Weibo (605 million monthly active users, as of 2023.)
- Threads (decentralized social media platform by Meta's Instagram, released in July 2023; 141 million users now.)





# Online Social Media

**Social media** has become a convenient way for people to exchange information.

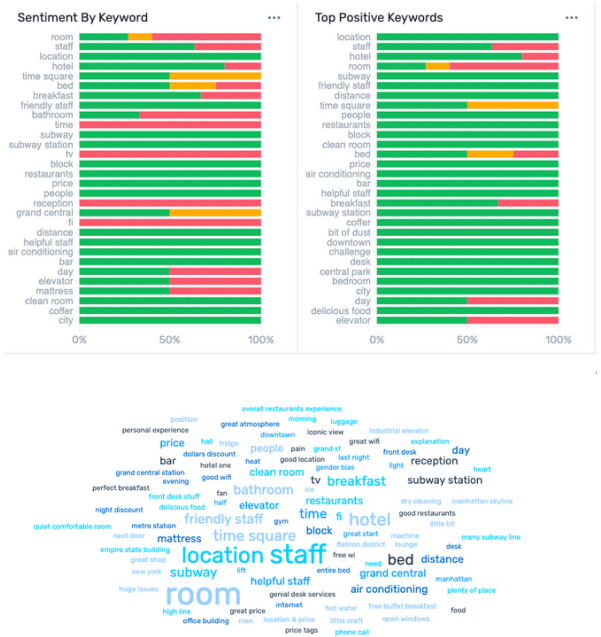
- Twitter (rebranded to X after Elon Musk's acquisition in 2023.)
- Weibo (605 million monthly active users, as of 2023.)
- Threads (decentralized social media platform by Meta's Instagram, released in July 2023; 141 million users now.)




## Why is social media important?

# Social Media as Information Sources

- Social media has become sources for information. For examples, for companies to understand **customers' opinions**, for investors to capture **markets' sentiment**. It also provides valuable insights during **elections**.



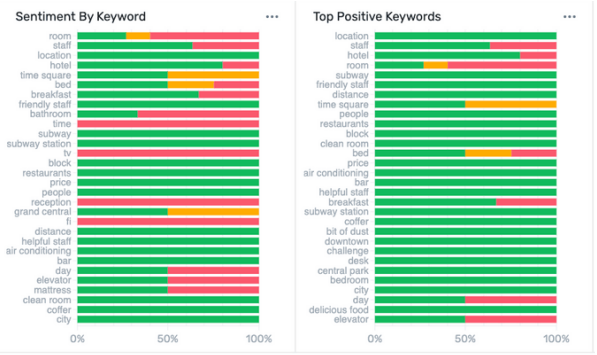

**(a) Brand Monitoring**


**(b) Social Velocity and Stock Market Tracing In Bloomberg**


**(c) Social Monitoring in Taiwan Election 2024**

# Social Media as Information Sources

- Social media has become sources for information. For examples, for companies to understand **customers' opinions**, for investors to capture **markets' sentiment**. It also provides valuable insights during **elections**.






Is social media a reliable source of information?

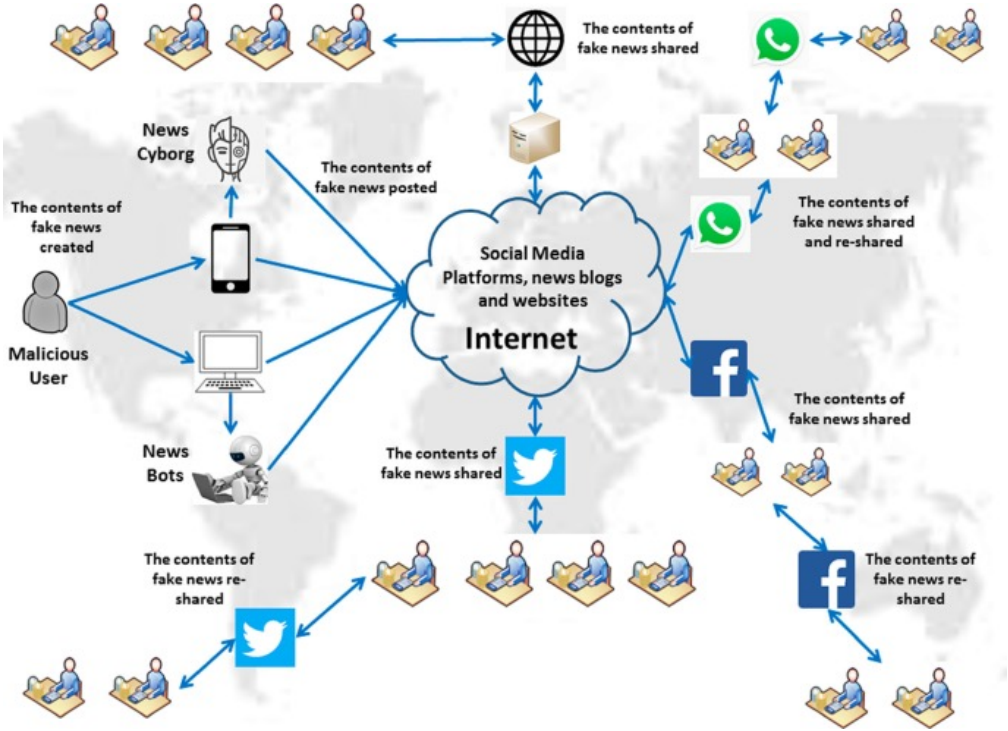
**(a) Brand Monitoring**

**(b) Social Velocity and Stock Market Tracing In Bloomberg**

**(c) Social Monitoring in Taiwan Election 2024**

# Misinformation on Social Media

- Social media provides a **real-time** platform for the **dissemination of information**, including the spread of fake news.






[1] Hakak et al., “Propagation of Fake News on Social Media: Challenges and Opportunities,” *Computational Data and Social Networks*, 2020.

# Manual Fact-Checking Websites

- **Manual fact-checking websites** provide platforms for people to verify statements, but they often rely on human annotations.


(a) PolitiFact


(b) HKBU FACT CHECK

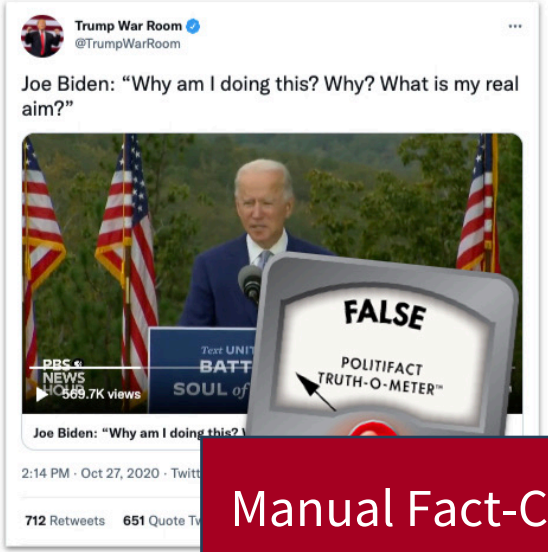

(c) Taiwan FactCheck Center



Examples of manual fact-checking websites.

# Manual Fact-Checking Websites

- **Manual fact-checking websites** provide platforms for people to verify statements, but they often rely on human annotations.






Manual Fact-Checking is time consuming and labor intensive.

**(a) PolitiFact** **(b) HKBU FACT CHECK** **(c) Taiwan FactCheck Center**

Examples of manual fact-checking websites.

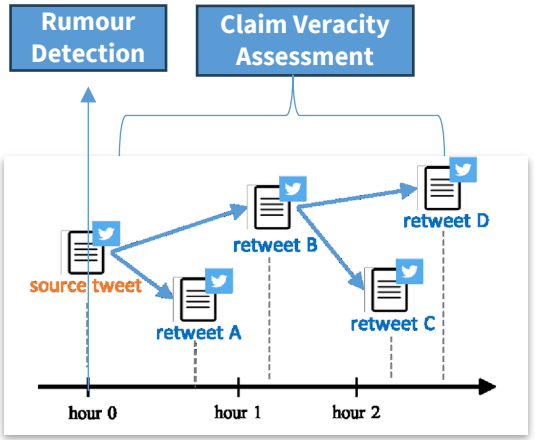
# Rumour Detection and Claim Veracity Assessment

## Task 1: Rumour Detection

- Detect unverified and false information (**source tweet**) **at posting**.
  - **Rumour**: Information unverified, of uncertain origin, or false at the time of posting.
  - **Non-Rumour**: Information verified and confirmed as true or not a rumour at the time of posting.

## Task 2: Claim Veracity Assessment

- Assess the truthfulness of rumours **post-posting**.
  - **True Rumour**: Information verified and confirmed as true.
  - **False Rumour**: Information verified and confirmed as false.
  - **Unverified Rumour**: Information for which the veracity cannot be confirmed.






[2] Zubiaga et al., “Detection and Resolution of Rumours in Social Media,” *ACM Computing Surveys*, vol. 51, no. 2, pp. 1–36, 2018.

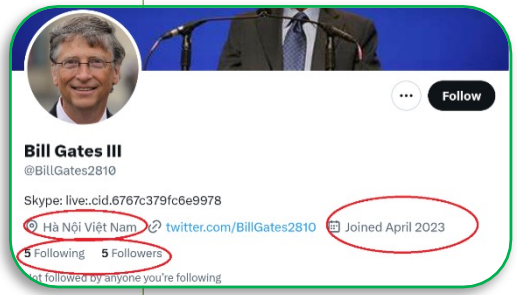
# Task 1: Rumour Detection

## I. Multimodal Source [3]

- **Diverse Content Types**: Text and images pose challenges (e.g., Instagram and Xiaohongshu).
- **Semantic Gap**: Bridging the semantic gap between textual and visual features is a key challenge in capturing nuanced relationships.


## II. User Credibility [4]

- **Dynamic User Behavior**: Adapting to evolving user behavior adds complexity to credibility assessment.
- **Contextual User Profiling**: Creating generalized models for user credibility requires accounting for contextual variations.

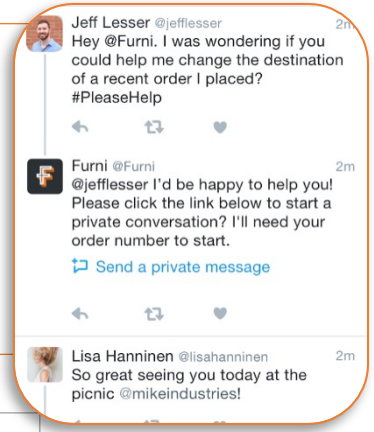



[3] T. Cheung and K. Lam, "Crossmodal bipolar attention for multimodal classification on social media," *Neurocomputing*, vol. 514, pp. 1–12, Dec. 2022.
[4] T. Cheung and K. Lam, "Author-Aware Rumour Detection with Layer-Wise Parameter-Efficient Tuning and Incomplete Feature Learning," submitted to *IEEE Access*.

# Task 2: Claim Veracity Assessment

## III. Community Response [5]

- **Diverse Reactions**: The wide array of responses to claims on social media introduces challenges in distinguishing between genuine and deceptive information.
- **Temporal Dynamics**: The evolving nature of community responses over time necessitates real-time analysis for accurate veracity assessment.




## IV. External Evidence [6]

- **Data Source Reliability**: Assessing the reliability of external evidence from various sources is a persistent challenge.
- **Integration Complexity:** Effectively combining and integrating external evidence demands sophisticated model architectures.


[5] T. Cheung and K. Lam, "Causal diffused graph-transformer network with stacked early classification loss for efficient stream classification of rumours," *Knowledge-Based Systems*, vol. 277, pp. 110807, 2023.
[6] T. Cheung and K. Lam, "FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking," in *Proceedings, APSIPA ASC 2023*, Oct. 31, 2023.

# Overview of This Thesis

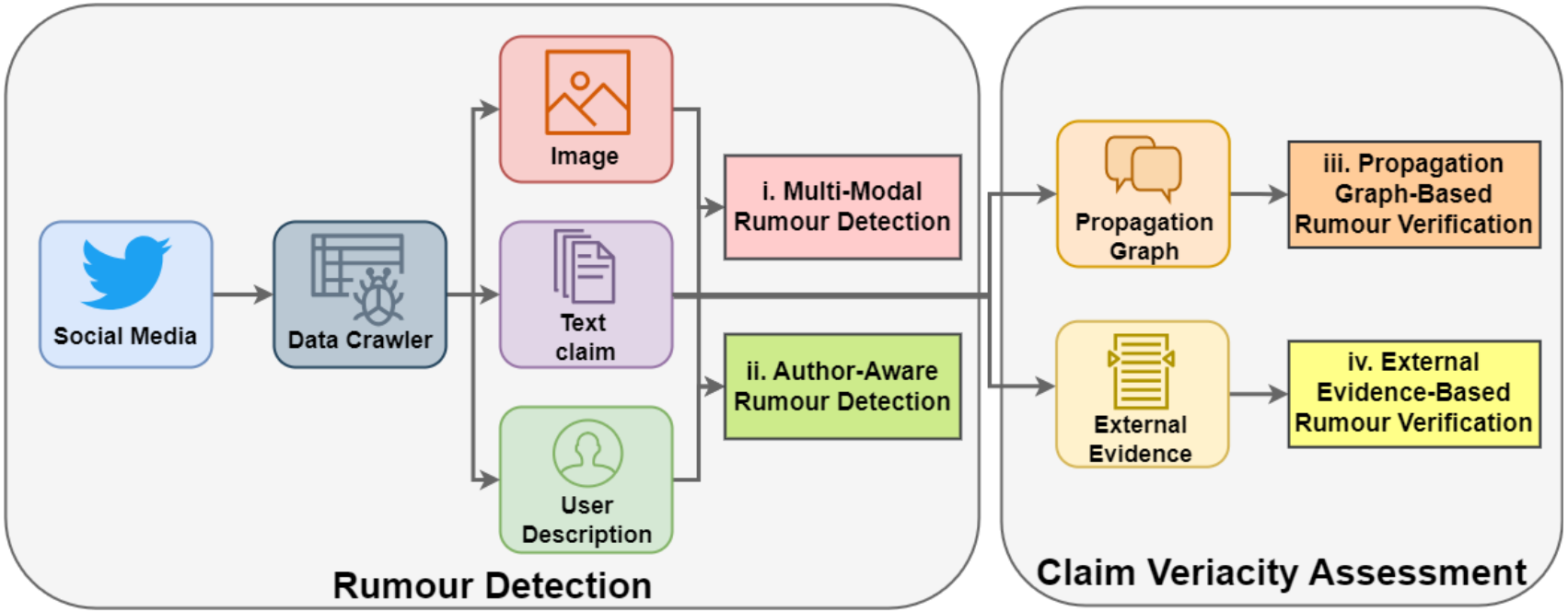



**Figure 3** An overview of the rumour detection and claim veracity assessment system on social media studied in this thesis.

# I. Multimodal Image-Text Rumour Detection

[3] T. Cheung and K. Lam, “Crossmodal bipolar attention for multimodal classification on social media,” *Neurocomputing*, vol. 514, pp. 1–12, Dec. 2022.

# Motivation

## Background

- Multimodal image-text rumor detection on social media has received increasing attention in recent years.


## Question

- How can we effectively model the semantic relationship between the two modalities?

## Challenge

- Existing methods ignore the inconsistent between the modalities. Both consistent and inconsistent semantic meanings between the modalities are important for multimodal classification on social media.

# Contributions

## Modeling Cross-Modal Relationships

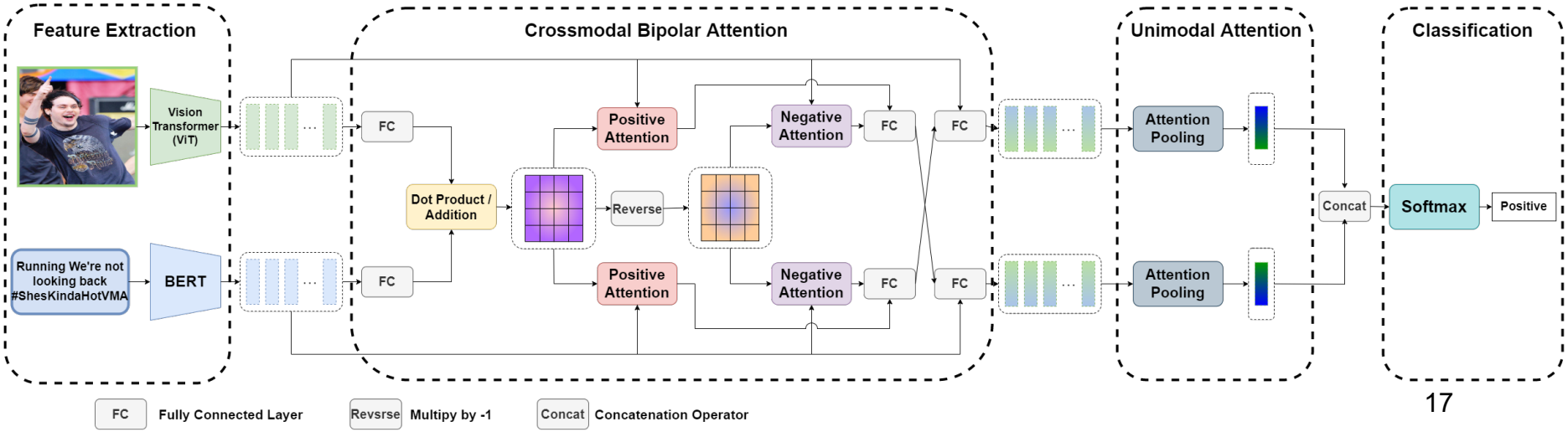
- A novel cross-modal bipolar attention (CBAN) mechanism.
- Direct and inverse textual-visual feature relationships.

## Feature Utilization and Aggregation

- Utilizing Vision Transformer (ViT)'s hidden representations for multimodal classification.
- Attentive pooling module for feature transformation.

# Model Overview of Crossmodal Bipolar Attention

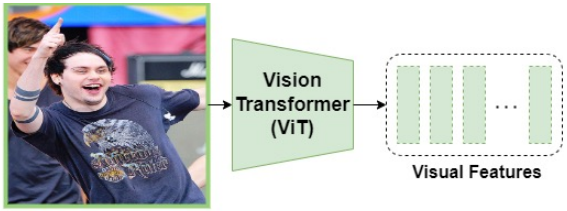
1. **Feature Extraction**: Encode the unimodal information into embeddings.
2. **Crossmodal Bipolar Attention**: Fuse the unimodal features into the cross-modal features.
3. **Unimodal Attention**: Aggregate the fused features into the final representation.
4. **Classification**: Classify the final representation into either rumour or non-rumour.



# Feature Extraction Modules

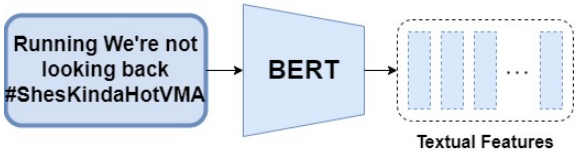
## Image Branch

- Vision Transformer (ViT) [7]
- The image is represented as $m$ visual features $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m]^{\mathrm{T}} \in \mathbb{R}^{m \times d_v}$.
  - $d_v$ is the embedding dimension of ViT.
  - $m$ is the number of region tokens in the input image.




## Text Branch

- Bidirectional Encoder Representations from Transformers (BERT) [8].
- The text is represented as $n$ textual vectors $\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d_t}$.
  - $d_t$ is the embedding dimension of BERT.
  - $n$ is the number of word tokens in the input sentence.






[7] Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ICLR*, 2021.
[8] Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL*, 2019.

# Crossmodal Bipolar Attention

## Fine-grained Similarity Matrix

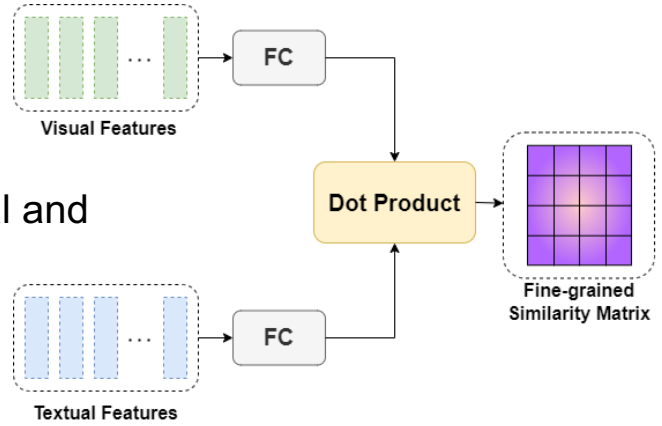
- First, align the textual features $\boldsymbol{T}$ and visual features $\boldsymbol{V}$ into the same shared embedding space, using two separate fully connected layers.

$$\mathbf{T}_{\mathbf{emb}} = \tanh(\mathbf{T}\mathbf{W}_{\mathrm{te}} + \mathbf{b}_{\mathrm{te}}),$$

$$\mathbf{V}_{\mathbf{emb}} = \tanh(\mathbf{V}\mathbf{W}_{\mathrm{ve}} + \mathbf{b}_{\mathrm{ve}}).$$

- The fine-grained similarity matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$ between every textual and visual feature is computed as:

$$\mathbf{S} = \frac{\mathbf{V}_{\mathbf{emb}}\mathbf{T}_{\mathbf{emb}}^{\mathrm{T}}}{\sqrt{d_e}}.$$



# Crossmodal Bipolar Attention

## Positive Attention Mechanism

- The visually guided textual features $\boldsymbol{T^p} \in \mathbb{R}^{m \times d_t}$ and the textually guided visual features $\boldsymbol{V^p} \in \mathbb{R}^{n \times d_v}$ are computed, as follows:

$$\boldsymbol{T^p} = \mathrm{softmax}(\boldsymbol{S})\boldsymbol{T},$$

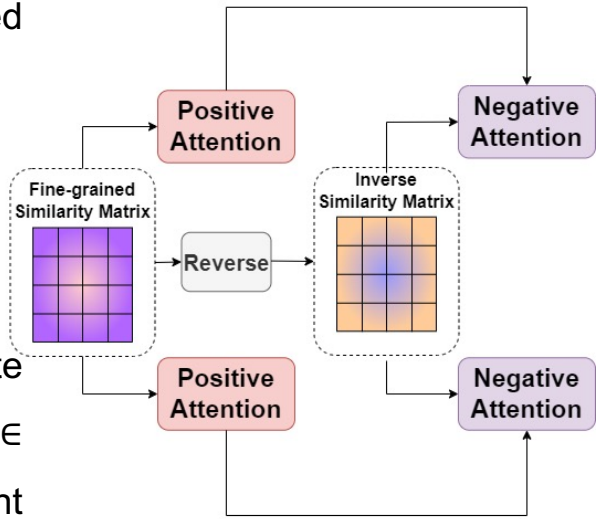
$$\boldsymbol{V^p} = \mathrm{softmax}\left(\boldsymbol{S}^{\mathrm{T}}\right)\boldsymbol{V}.$$

## Negative Attention Mechanism

- In addition to the positive-correlated attention vectors, we further compute the negatively correlated attention vectors, i.e., $\boldsymbol{T^n} \in \mathbb{R}^{m \times d_t}$ and $\boldsymbol{V^n} \in \mathbb{R}^{n \times d_v}$, by multiplying the similarity matrix $\boldsymbol{S}$ with a negative constant before applying Softmax, as follows:

$$\boldsymbol{T^n} = \mathrm{softmax}(-\boldsymbol{S})\boldsymbol{T},$$

$$\boldsymbol{V^n} = \mathrm{softmax}\left(-\boldsymbol{S}^{\mathrm{T}}\right)\boldsymbol{V}.$$



# Attention Vector Aggregation

## Attention Vector Concatenation

- A fully connected layer is used to combine the positive and negative attention vectors:

$$\boldsymbol{T}^* = \tanh\big((\boldsymbol{T}^p \oplus \boldsymbol{T}^n)\boldsymbol{W}_{tt} + \boldsymbol{b}_{tt}\big),$$

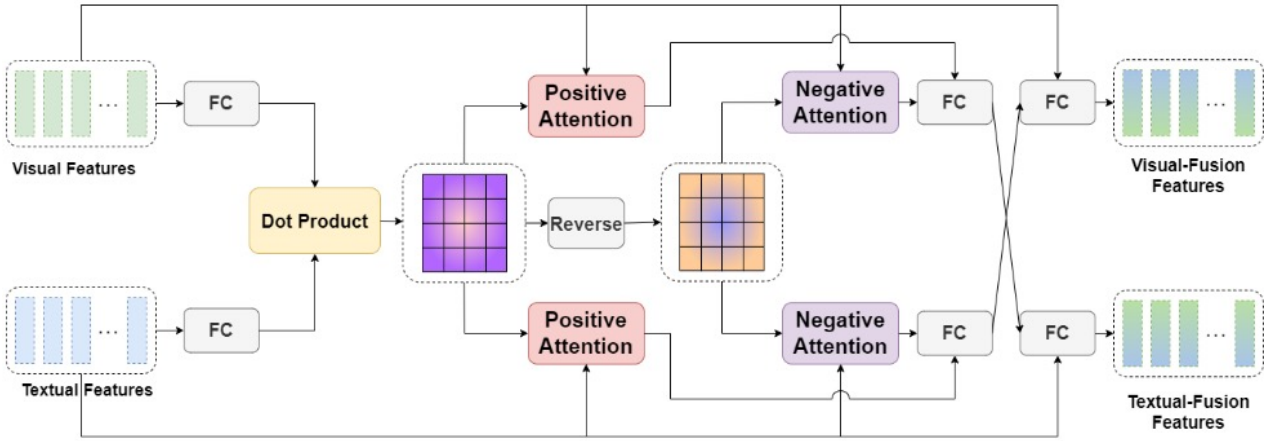
$$\boldsymbol{V}^* = \tanh\big((\boldsymbol{V}^p \oplus \boldsymbol{V}^n)\boldsymbol{W}_{vv} + \boldsymbol{b}_{vv}\big).$$

## Embedding Vector Concatenation

- To keep the semantic information of the embedding, a fully connected layer is used to combine the embedding and attention vectors:

$$\boldsymbol{V}_f = \tanh\big((\boldsymbol{V} \oplus \boldsymbol{T}^*)\boldsymbol{W}_v + \boldsymbol{b}_v\big),$$

$$\boldsymbol{T}_f = \tanh\big((\boldsymbol{T} \oplus \boldsymbol{V}^*)\boldsymbol{W}_t + \boldsymbol{b}_t\big).$$



# Unimodal Attention Pooling

## Aggerated vision fusion features

- $\boldsymbol{V}_f$ attention score $\alpha_i^v$ for each feature $\boldsymbol{v}_{f,i}$ as follows:

$$\alpha_i^v = \frac{\boldsymbol{v}_{f,i}\boldsymbol{U}_d}{\sqrt{d_v}},$$

$$\tilde{\alpha}_i^v = \frac{\exp(\alpha_i^v)}{\sum_{i=1}^{m}\exp(\alpha_i^v)},$$

$$\boldsymbol{v}_f' = \sum_{i=1}^{m}\tilde{\alpha}_i^v\,\boldsymbol{v}_{f,i}.$$

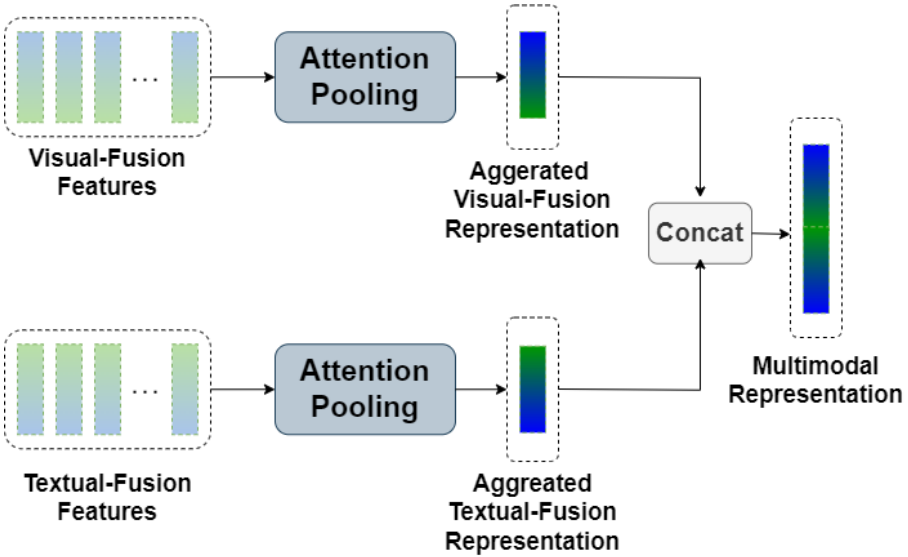
## Aggerated textual fusion features

- $\boldsymbol{T}_f$ attention score $\alpha_i^t$ for each feature $\boldsymbol{t}_{f,i}$ as follows:

$$\alpha_i^v = \frac{\boldsymbol{t}_{f,i}\boldsymbol{U}_d}{\sqrt{d_t}},$$

$$\tilde{\alpha}_i^t = \frac{\exp(\alpha_i^t)}{\sum_{i=1}^{n}\exp(\alpha_i^t)},$$

$$\boldsymbol{t}_f' = \sum_{i=1}^{n}\tilde{\alpha}_i^v\,\boldsymbol{v}_{t,i}.$$



# Classification Layer and Loss Function

## Classification Layer

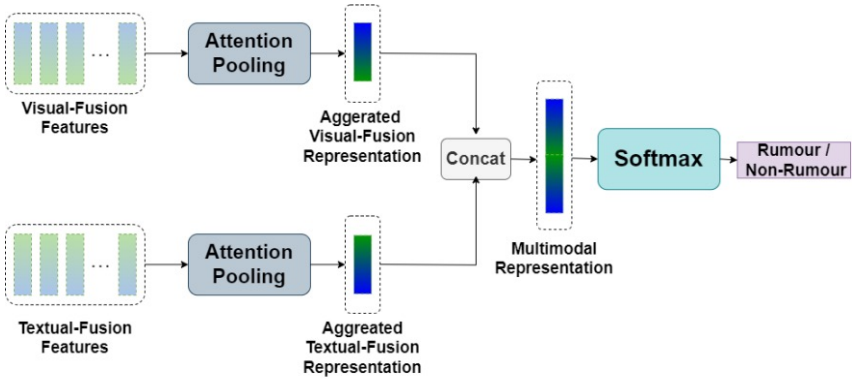
- The two final fused features, i.e., $\boldsymbol{v}_f'$ and $\boldsymbol{t}_f'$, are used to perform classification. The concatenated feature vector is classified by using a fully connected layer, as follows:

$$\hat{\boldsymbol{y}} = \tanh\left((\boldsymbol{v}_f' \oplus \boldsymbol{t}_f')\boldsymbol{W}_c + \boldsymbol{b}_c\right),$$

## Loss Function

- The cross-entropy loss is used as the objective function. Given the predicted label $\hat{\boldsymbol{y}}$ and the ground-truth label $\boldsymbol{y}$, we minimize the negative log-likelihood with $c$ classes after the Softmax function, as follows:

$$Loss = -\sum_i^c \boldsymbol{y}\log\left(\text{softmax}(\hat{\boldsymbol{y}})\right).$$



# Experimental Details

## Datasets

- PHEME [9]: 1972 rumours, 3030 non-rumours, from Twitter.
- Weibo [10]: 6226 rumours, 9405 non-rumours, from Weibo.

## Hyperparameters

- For all experiments, the models were trained with a mini-batch size of 64 for 10 epochs. We use the Adam Optimizer with a fixed learning rate of 0.00002.

[9] Augbiaga et al., “Exploiting Context for Rumour Detection in Social Media,” *Int. Conf. on Social Informatics*,” 2017.
[10] Jin et al., “Multimodal Fusion with Recurrent Neural Networks for Rumour Detection on Microblogs,” *ACM MM*, 2017.

# Results on PHEME Dataset

| Methods | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| SVM [42] | 0.639 | 0.638 | 0.641 | 0.639 |
| GRU [28] | 0.832 | 0.819 | 0.804 | 0.805 |
| CNN [43] | 0.779 | 0.766 | 0.741 | 0.749 |
| TextGCN [44] | 0.828 | 0.801 | 0.782 | 0.783 |
| Att-RNN [18] | 0.850 | 0.834 | 0.824 | 0.829 |
| EANN [16] | 0.681 | 0.693 | 0.707 | 0.721 |
| MVAE [45] | 0.852 | 0.839 | 0.818 | 0.827 |
| SAFE [46] | 0.811 | 0.817 | 0.750 | 0.767 |
| MMCN [47] | 0.872 | 0.863 | 0.850 | 0.856 |
| **CBAN (Ours)** | **0.894** | **0.868** | **0.878** | **0.894** |

# Results on WEIBO Dataset

| Methods | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| SVM [42] | 0.640 | 0.696 | 0.686 | 0.679 |
| GRU [28] | 0.720 | 0.709 | 0.702 | 0.699 |
| CNN [43] | 0.740 | 0.742 | 0.740 | 0.740 |
| TextGCN [44] | 0.787 | 0.844 | 0.863 | 0.777 |
| Att-RNN [18] | 0.772 | 0.787 | 0.838 | 0.769 |
| EANN [16] | 0.782 | 0.790 | 0.818 | 0.780 |
| MVAE [45] | 0.824 | 0.828 | 0.829 | 0.823 |
| SAFE [46] | 0.763 | 0.775 | 0.846 | 0.761 |
| MMCN [47] | 0.879 | 0.880 | 0.880 | 0.880 |
| **CBAN (Ours)** | **0.934** | **0.935** | **0.934** | **0.934** |

# Ablation Studies on CBAN

| Dataset | Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| PHEME | CBAN w/o crossmodal attention | 0.885 | 0.854 | 0.883 | 0.866 |
| | CBAN w/o unimodal attention | 0.883 | 0.854 | 0.868 | 0.860 |
| | **CBAN (Full)** | **0.894** | **0.868** | **0.878** | **0.894** |
| Weibo | CBAN w/o crossmodal attention | 0.881 | 0.882 | 0.881 | 0.881 |
| | CBAN w/o unimodal attention | 0.918 | 0.921 | 0.918 | 0.918 |
| | **CBAN (Full)** | **0.934** | **0.935** | **0.934** | **0.934** |

# Examples of Multimodal Rumours and Non-Rumours

| Image | Text | Label |
|---|---|---|
|  | BREAKING: Germanwings airplane crashes in French Alps with 142 passengers. | Rumour |
|  | call me foolish but this girl holding #JeSuisCharlie hiding her face, in a ghostly looking #Aleppo #Syria made me cry | Non-Rumour |

# Examples of Multimodal Rumours and Non-Rumours

| Image | Text | Label |
|---|---|---|
|  | 紧急通知！紧急通知！经辽宁省地震局公布，于2015年8月4日，15：30时左右鞍山，海城，盘锦，营口地区会出现5.8级地震<br>(emegency notice! emegency notice! According to the Liaoning Provincial Seismological Bureau, an earthquake with a magnitude of 5.8 will occur in Anshan, Haicheng, Panjin, and Yingkou areas around 15:30 on August 4, 2015.) | Rumour |
|  | 今早，暴雨、防汛防台均橙色预警，目前，闵行、青浦、普陀等多处已“见海”。<br>(This morning, orange warnings were issued for heavy rain and flood control. At present, the sea has already appeared in Minhang, Qingpu, Putuo and other places.) | Non-Rumour |

# II. User-Aware Rumour Detection

[4] T. Cheung and K. Lam, “Author-Aware Rumour Detection with Layer-Wise Parameter-Efficient Tuning and Incomplete Feature Learning,” submitted to *IEEE Access*.

# Challenges in Existing Rumor Detection Methods

## Challenges

- Current rumour detection methods often hinge on insights from user comments and other social media metrics.
- Delay in the detection process arise due to the lagging nature of crowd signals.

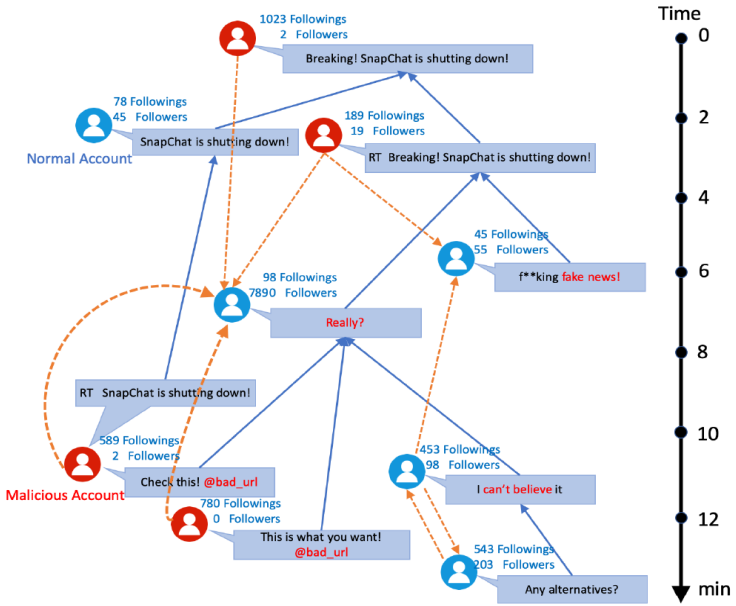
## Solution

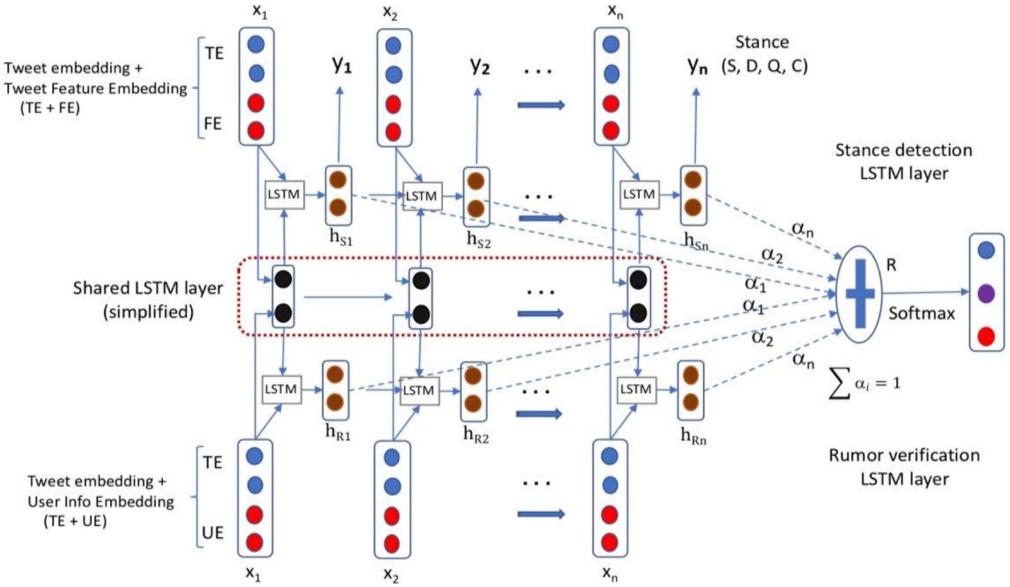
- User profiling provides useful credibility signal to distinguish malicious behaviors, such as spreading misinformation on social media.


# Limitation in Existing Author-Aware Approches

## Inefficiency in utilizing Pretrained Models

- Delayed crowed signal in utilizing social network relationship.

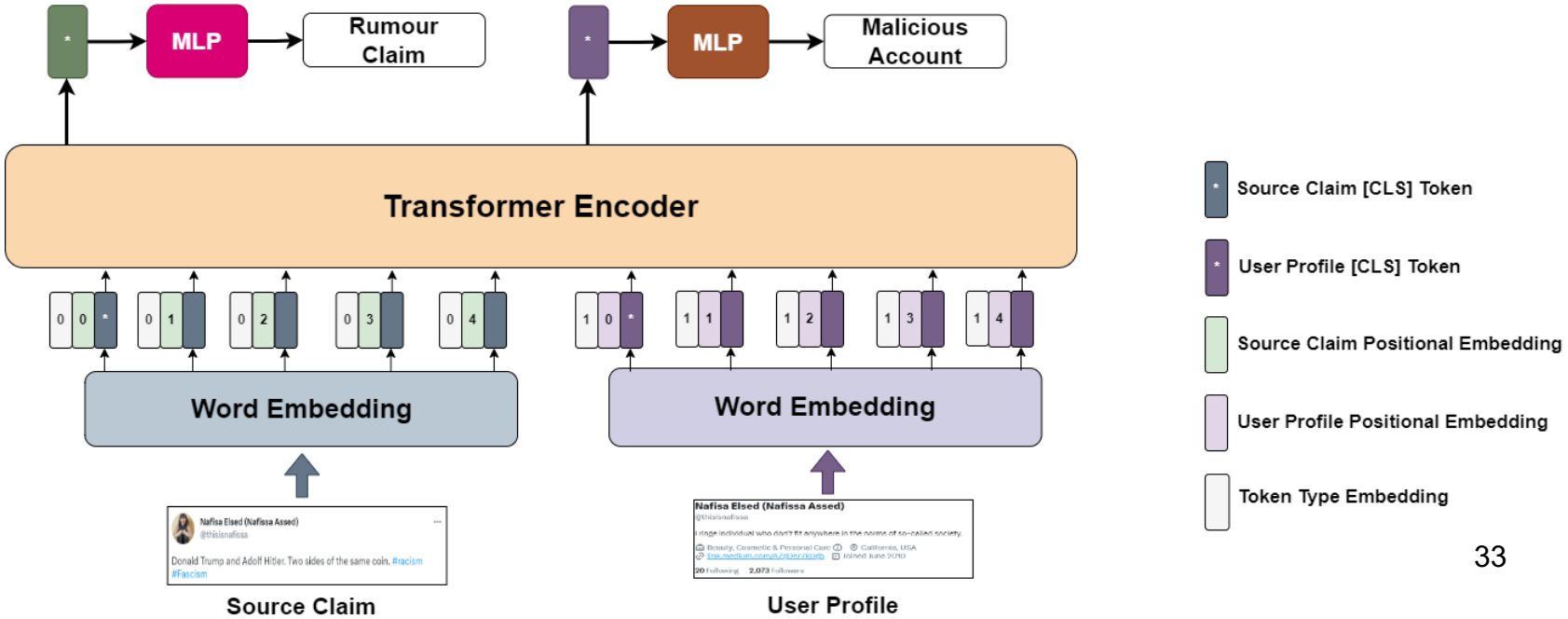


# Author-Aware Rumour Detection Framework

## Tuning Pretrained Transformer Encoder for Author-Aware Rumour Detection

- Our goal is to fine-tune a Transformer encoder that accepts both source claim and user profile, for author-aware rumour detection.



# Tokenization and Input Embeddings

## Source Claim Features

- The source claim feature $\overline{\boldsymbol{T}} \in \mathbb{R}^{m \times d_e}$ is the addition of text embedding $\boldsymbol{T}_{\mathbf{emb}}$, position embedding $\boldsymbol{T}_{\boldsymbol{pos}}$ and type embedding $\boldsymbol{T}_{type}$.

$$\overline{\boldsymbol{T}} = \boldsymbol{T}_{\boldsymbol{emb}} + \boldsymbol{T}_{\boldsymbol{pos}} + \boldsymbol{T}_{\boldsymbol{type}}.$$

## User Profile Features

- The user profile feature $\overline{\boldsymbol{U}} \in \mathbb{R}^{n \times d_e}$ is the addition of user embedding $\boldsymbol{U}_{\boldsymbol{emb}}$, position embedding $\boldsymbol{U}_{\boldsymbol{pos}}$ and type embedding $\boldsymbol{U}_{\boldsymbol{type}}$.

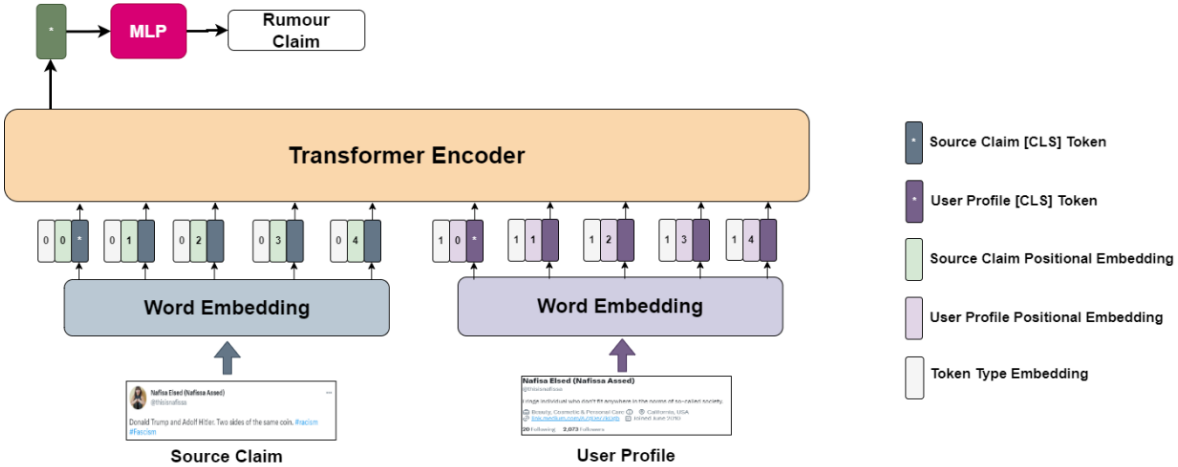
$$\overline{\boldsymbol{U}} = \boldsymbol{U}_{\boldsymbol{emb}} + \boldsymbol{U}_{\boldsymbol{pos}} + \boldsymbol{U}_{\boldsymbol{type}}.$$

## Author-Aware Rumour Detection

- Concatenate the text and user features to obtain the fine-grained features $\overline{\boldsymbol{F}}$, before sending to the Transformer models, as follows:

$$\overline{\boldsymbol{F}} = [\overline{\boldsymbol{T}}, \overline{\boldsymbol{U}}],$$

$$\boldsymbol{f_{cls}} = \text{Transformer}(\overline{\boldsymbol{F}}).$$



# Multi-Task Transformer with Missing Features

## Rumour Claim Detection

- To classify rumour claims without author profiles, we send the text features $\overline{\boldsymbol{T}}$ into the transformer-based pretrained language models, to obtain the [CLS] token of output source claim representation $\mathbf{t_{cls}} \in \mathbb{R}^{1\times d_e}$, as follows:

$$\mathbf{t_{cls}} = \text{Transformer}_{\text{text}}(\overline{\boldsymbol{T}}).$$

## Malicious User Detection

- To classify malicious users, we send the user features $\overline{\boldsymbol{U}}$ into the transformer model, to obtain the [CLS] token of output user representation $\mathbf{u_{cls}} \in \mathbb{R}^{1\times d_e}$, as follows:

$$\mathbf{u_{cls}} = \text{Transformer}_{\text{user}}(\overline{\boldsymbol{U}}).$$

# Multi-Task Loss with Missing Features

## Multi-Task Learning with Missing Features

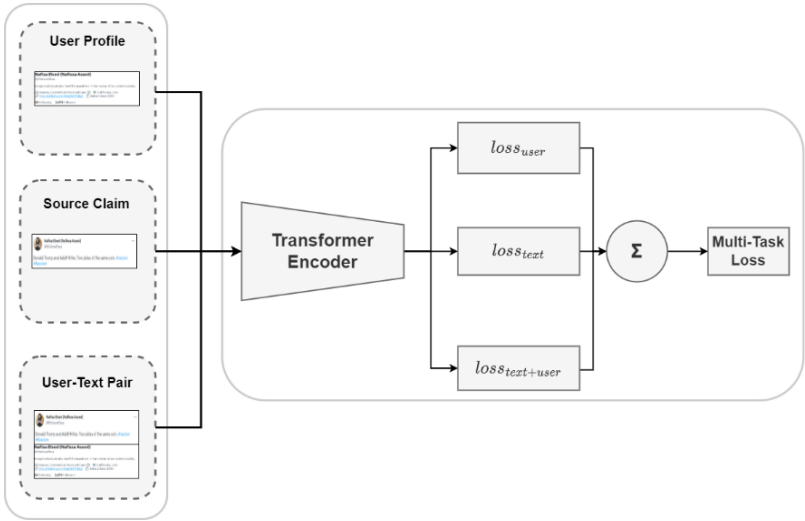
- Softmax classifier is used to predict whether it is a rumour or not, as follows:

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{W_c f_{cls}} + \boldsymbol{b_c}),$$

$$\text{loss} = -\big(\boldsymbol{y}\log(\hat{\boldsymbol{y}}) + (\boldsymbol{1} - \boldsymbol{y})\log(1 - \hat{\boldsymbol{y}})\big).$$

- The multitask learning loss with missing features (i.e., rumour claim detection, malicious account detection, and author-aware rumour detection) as follows:

$$\text{loss}_{\text{Total}} = \text{loss}_{\text{text}} + \text{loss}_{\text{user}} + \text{loss}_{\text{multi}}.$$



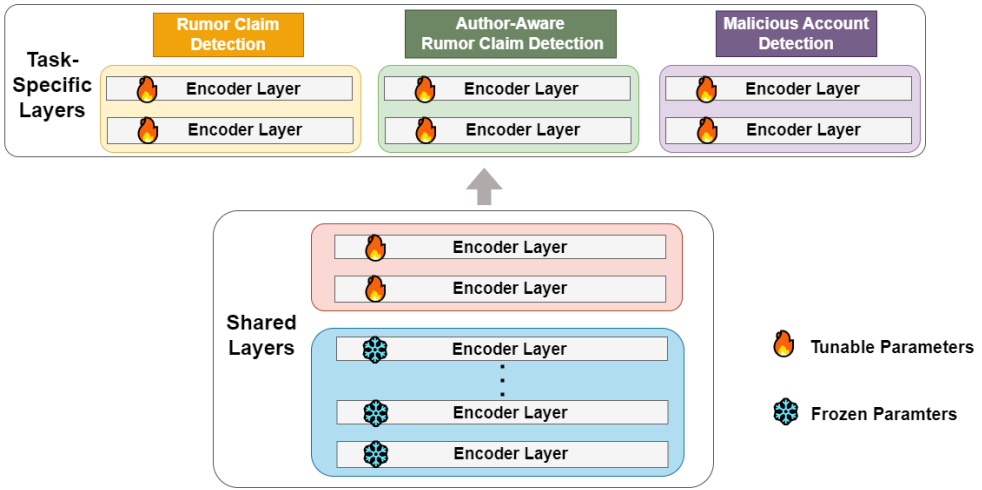
# Layer-Wise Parameter-Efficient Tuning (LWPEF)

## Transfer Learning

- Layers from Layer 1 to Layer $b$ are frozen, allowing shared but unaltered parameters during tuning.

## Hard-Parameter Sharing

- Layers from Layer $b$ to Layer $c-1$ are shared and updated across tasks during tuning.
- Layers from Layer $c$ to the final layer are task-specific layers learnt for each task.



# Experimental Details

## Datasets

- Twitter15 [11] Dataset: 374 rumours, 1116 non-rumours, from Twitter.
- Twitter16 [11] Dataset: 205 rumours, 613 non-rumours, from Twitter.
- CR-Twitter [12] Dataset: 3616 rumours, 6295 non-rumours, from Twitter.

## Hyperparameters

- We trained the models for 30 epochs with a mini-batch size of 16, using the Adam optimizer with a learning rate of 0.00002.
- For the LWPET, we set the values $b = 8$, $c = 10$.

[11] Ma et al., “rumour Detection on Twitter with Tree-structured Recursive Neural Networks,” *ACL*,” 2018.
[12] Chen et al., “Identifying Cantonese rumors with discriminative feature integration in online social networks,” *Expert Systems with Applications*, 2020.

# Results on Twitter15, Twitter16, and CR-Twitter Datasets

| Dataset | Method | Author Profile Injection | Macro-F1 | Accuracy |
|---|---|---|---|---|
| Twitter15 | BERT | × | 0.8505 | 0.8814 |
| | | ✓ | **0.9413** | **0.9526** |
| | RoBERTa | × | 0.8985 | 0.9170 |
| | | ✓ | 0.9366 | 0.9486 |
| | DistilBERT | × | 0.8472 | 0.8735 |
| | | ✓ | 0.9372 | 0.9486 |
| | DeBERTa | × | 0.8958 | 0.9130 |
| | | ✓ | 0.9016 | 0.9170 |
| Twitter16 | BERT | × | 0.8328 | 0.8702 |
| | | ✓ | **0.9270** | **0.9389** |
| | RoBERTa | × | 0.8358 | 0.8702 |
| | | ✓ | 0.8888 | 0.9084 |
| | DistilBERT | × | 0.8440 | 0.8779 |
| | | ✓ | 0.9073 | 0.9237 |
| | DeBERTa | × | 0.8085 | 0.8473 |
| | | ✓ | 0.8786 | 0.9008 |
| CR-Twitter | BERT | × | 0.7996 | 0.8134 |
| | | ✓ | **0.8302** | **0.8411** |
| | RoBERTa | × | 0.7927 | 0.8103 |
| | | ✓ | 0.8174 | 0.8282 |

# Ablation studies on Author-Aware Rumour Detection

| Dataset | Model | Macro-F1 | Accuracy |
|---|---|---|---|
| **Twitter-15** | Ours w/o user profiling | 0.8505 | 0.8814 |
| | Ours w/o multi-task learning | 0.9236 | 0.9368 |
| | **Ours (Full)** | **0.9413** | **0.9526** |
| **Twitter-16** | Ours w/o user profiling | 0.8328 | 0.8702 |
| | Ours w/o multi-task learning | 0.8988 | 0.9160 |
| | **Ours (Full)** | **0.9270** | **0.9389** |
| **CR-Twitter** | Ours w/o user profiling | 0.7996 | 0.8134 |
| | Ours w/o multi-task learning | 0.8162 | 0.8288 |
| | **Ours (Full)** | **0.8302** | **0.8411** |

# Different Fusion Strategies

| Dataset | Fusion Strategies | Macro-F1 | Accuracy |
|---|---|---|---|
| **Twitter-15** | Feature concatenation in output feature | 0.8968 | 0.9170 |
| | Score aggregation in output logit | 0.8737 | 0.9012 |
| | **Token concatenation in input embedding** | **0.9236** | **0.9368** |
| **Twitter-16** | Feature concatenation in output feature | 0.8085 | 0.8473 |
| | Score aggregation in output logit | 0.8468 | 0.8779 |
| | **Token concatenation in input embedding** | **0.8988** | **0.9160** |
| **CR-Twitter** | Feature concatenation in the output feature | 0.7964 | 0.8159 |
| | Score aggregation in output logit | 0.7948 | 0.8159 |
| | **Token concatenation in input embedding** | **0.8162** | **0.8288** |

# Different Tuning Strategies

| Dataset | Tuning Strategy | Macro-F1 | Accuracy |
|---|---|---|---|
| **Twitter-15** | Fine-Tuning [41] | 0.9040 | 0.9249 |
| | Adapter-Tuning with Bottleneck [71] | 0.8753 | 0.8933 |
| | Adapter-Tuning with LORA [72] | 0.8376 | 0.8656 |
| | Prefix-Tuning [73] | 0.8623 | 0.8933 |
| | **Layer-Wise Parameter-Efficient Tuning (Ours)** | **0.9413** | **0.9526** |
| **Twitter-16** | Fine-Tuning [41] | 0.9270 | 0.9389 |
| | Adapter-Tuning with Bottleneck | 0.8641 | 0.8855 |
| | Adapter-Tuning with LORA [72] | 0.7694 | 0.8244 |
| | Prompt-Tuning [73] | 0.7585 | 0.8092 |
| | **Layer-Wise Parameter-Efficient Tuning (Ours)** | **0.9270** | **0.9389** |
| **CR-Twitter** | Fine-Tuning [41] | 0.8050 | 0.8153 |
| | Adapter-Tuning with Bottleneck [71] | 0.8017 | 0.8147 |
| | Adapter-Tuning with LORA [72] | 0.7858 | 0.8023 |
| | Prompt-Tuning [73] | 0.7912 | 0.8073 |
| | **Layer-Wise Parameter-Efficient Tuning (Ours)** | **0.8302** | **0.8411** |

# Examples of Rumour and Non-rumour

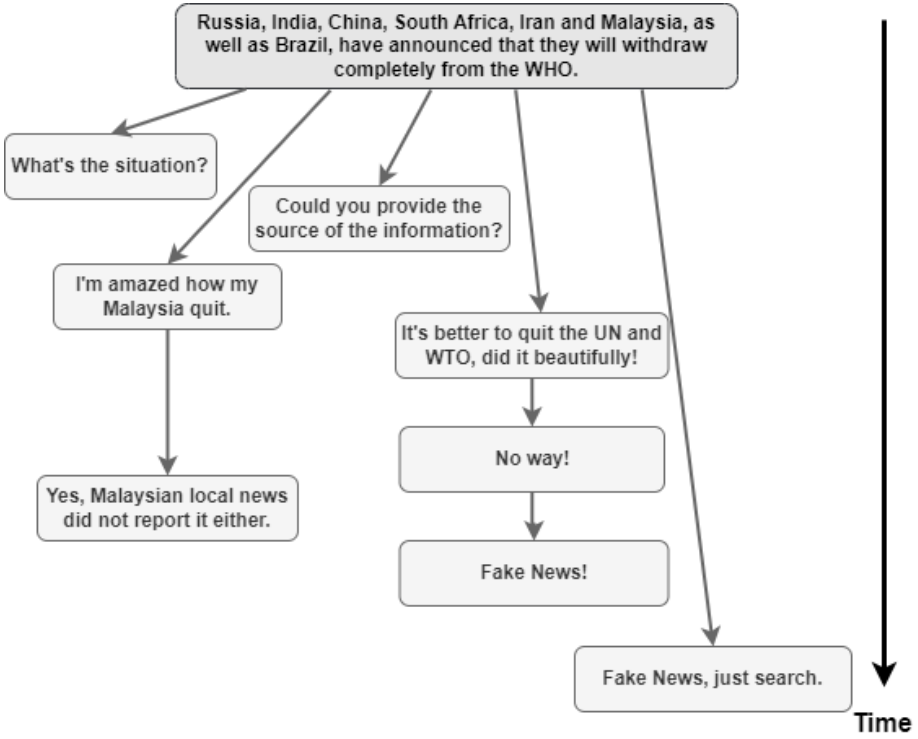
| | Text | User | Label |
|---|---|---|---|
| (a) | Hillary Clinton is lying. Clinton is NOT the first woman nominated for president. | Just google me. It's pretty funny. | Rumour |
| (b) | Doctors should screen all adults for depression at least once, the task force recommends. | Breaking news from CNN Digital. | Non-Rumour |
| (c) | 福建晉江某人隱瞞武漢旅行史，過年期間活躍出席當地各種公開活動和宴席，導致3-4千人需要被監控。<br>(A person in Jinjiang, Fujian concealed his travel history to Wuhan, and actively attended various local public events and banquets during the Chinese New Year, resulting in 3-4 thousand people needing to be monitored.) | RFA自由亞洲電台政治漫畫家/個人言論与RFA公司立場無關<br>(RFA Radio Free Asia political cartoonist/personal remarks have nothing to do with RFA's position) | Rumour |
| (d) | 林鄭月娥由上海轉往南京 ，出席第二屆蘇港融合發展峰會，並與江蘇省領導會面，又與港商交流。<br>(Carrie Lam transferred from Shanghai to Nanjing to attend the 2nd Suzhou-Hong Kong Integration Development Summit, met with leaders of Jiangsu Province, and communicated with Hong Kong businessmen.) | 香港特區政府網上新聞平台<br>(Hong Kong SAR Government Online News Platform) | Non-Rumour |

# III. Stream Verification of Online rumours using Community Response

[5] T. Cheung and K. Lam, “Causal diffused graph-transformer network with stacked early classification loss for efficient stream classification of rumours,” *Knowledge-Based Systems*, vol. 277, pp. 110807, 2023.

## Community Response

- Online social media platforms allow users to quote and reply to a source post or a reply. The propagation graph evolves along time.



# Limitation of Existing Approches

## Graph Neural Networks (GNNs) and Transformers

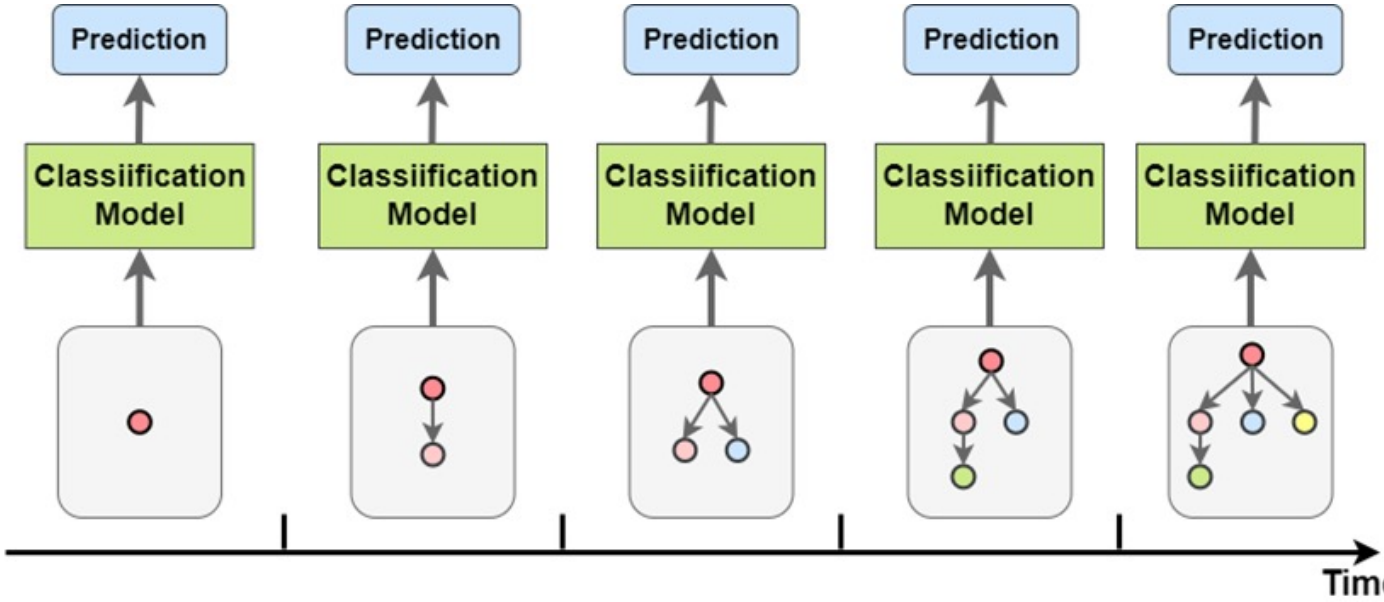
- GNNs model social media propagation patterns.
- Transformers capture sequential source-to-reply relationships.
- Both networks show promising rumour classification performance.

## Non-Causal Efficiency Challenge

- Non-causal nature hinders efficiency in stream mining.
- Features from a reply depend on future replies, impacting real-time verification.

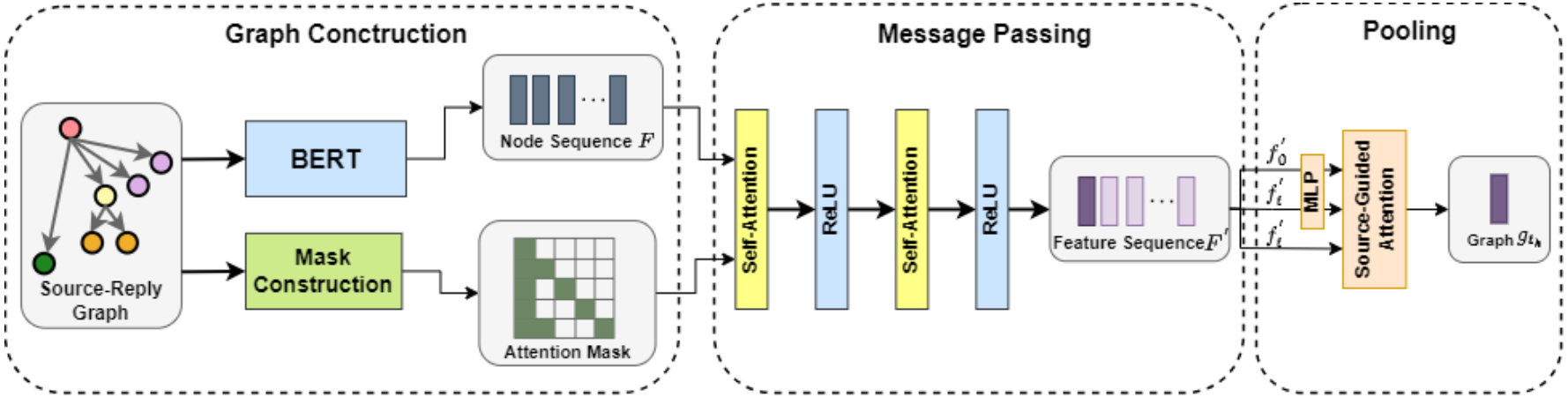
# Stream Verification of rumours

- **Early stream verification of rumours** is to instantly determine the veracity of the source post, whenever a reply is posted, favourably when the number of replies is small in the early stage of propagation.



# Causal Diffused Graph-Transformer Network (CDGTN)

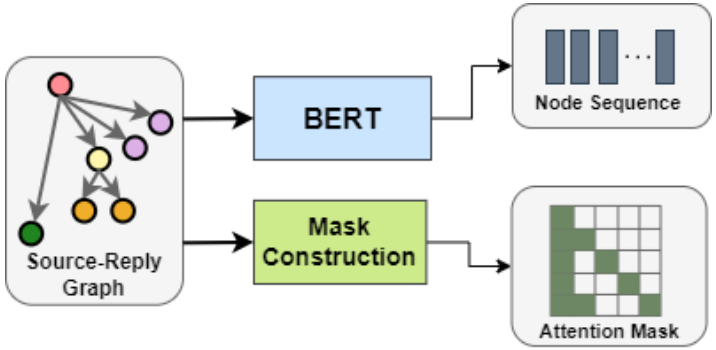
1. **Graph Construction**: Construct the source-reply graph by obtaining the node features and graph connectivity mask.
2. **Message Passing**: Aggregate information from its local neighborhood
3. **Pooling**: Aggregate the node features into the graph representation.



# Casual Dffused Graph-Transformer Network

## Node Feature Matrix

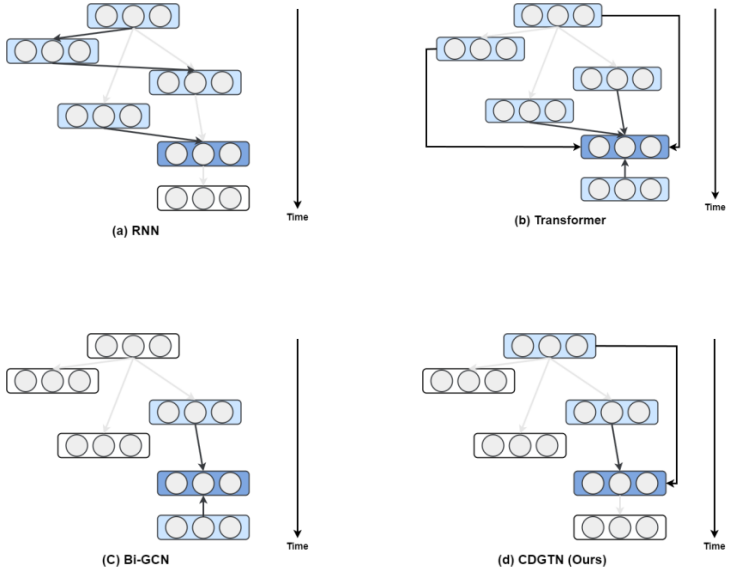
- BERT is used to project each tweet into a fix-length representation. The hidden representation of the [CLS] token is used as the representation for each tweet.




## Edge Connectivity Matrix

- Forces the query not to attend to the key and value vectors in future positions.

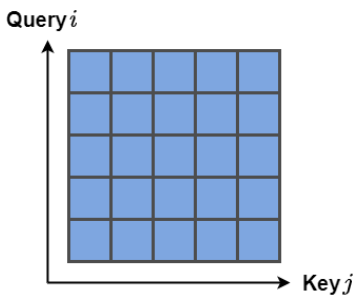
- Compute the attention scores that belong to the path of positions from the source post to the latest reply.



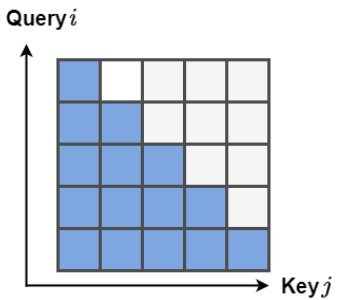
# Illustration of Attention Mask

## Causal Graph-Aware Attention Mask

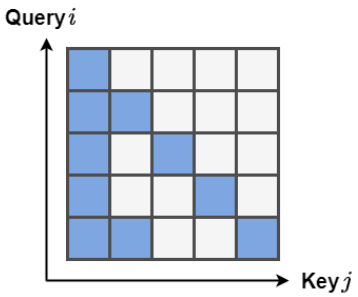
- Force the query not to attend to the key and value vectors in future positions.
- Compute the attention scores that belong to the path of positions from the source post to the latest reply.



(a) Full Attention



(b) Causal Attention

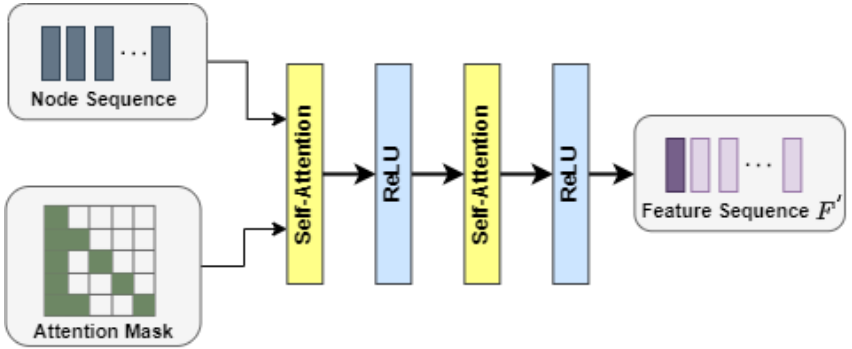


(c) Causal Graph-Aware Attention

# Message Passing and Causal Representations

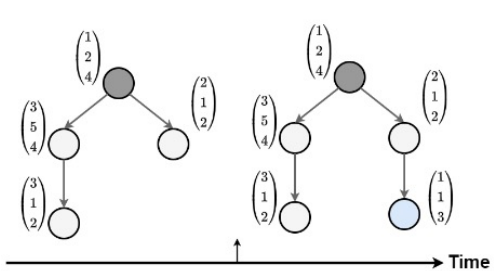
## Message Passing

- We use a two-layer Transformer encoder to model a message and the corresponding nodes all the way to the source post.

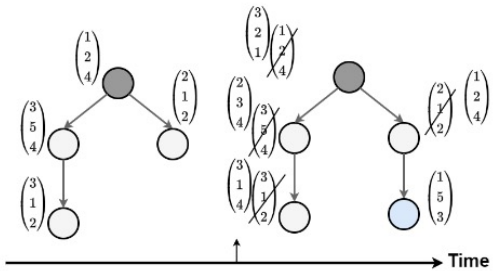



## Causal Hidden Representations

- Due to the nature of the proposed attention mask, the hidden representations of the graph is causal.




(a) Causal Hidden Representations




(b) Non-Causal Hidden Representations

## Source-Guided Incremental Attention Pooling

- We use Multilayer Perceptron (MLP) to calculate the importance of the feature $a_t$ at any timestamp $t$, given by:

$$a_t = \boldsymbol{W_2}\mathrm{Tanh}\,(\boldsymbol{W_1}(\,\boldsymbol{f'_0} \oplus \boldsymbol{f'_t}\,)),$$

$$\widehat{a_t} = \frac{e^{a_t}}{\sum_{t=1}^{t_k} e^{a_t}}.$$

- The final representation of a conversation at the final time $t_k$ is calculated by:

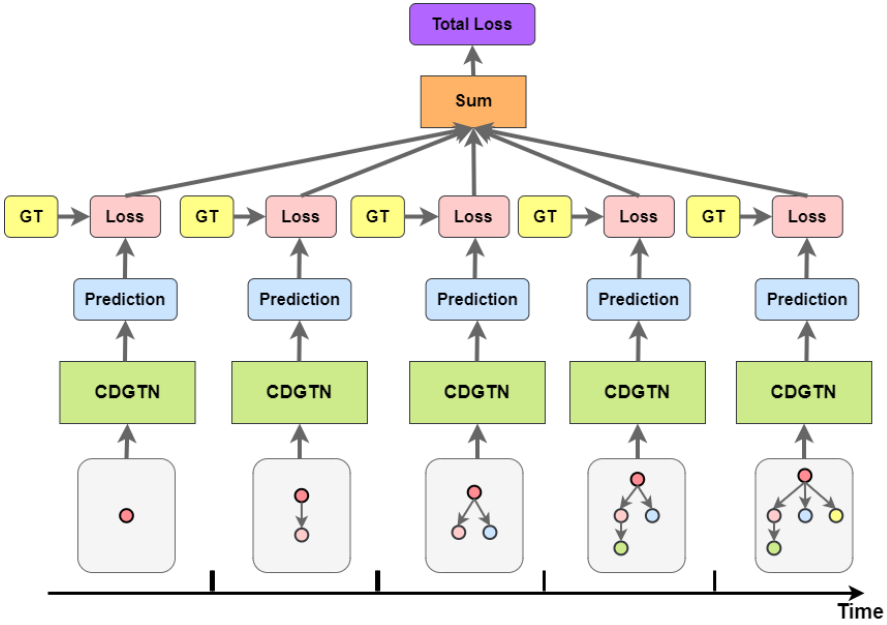
$$\boldsymbol{g_{t_k}} = \sum_{t=1}^{t_k} \widehat{a_t}\boldsymbol{f'_t}.$$

# Illustration of Stacked Early Classification Loss (SecLoss)

## Stacked Early Classification Loss

- Our goal is to learn the sum of errors from all predictions at any time $t$.
- Mathematically, we use the Softmax classification and minimize the summation of predictions at all timestamps $t \in \{0,1,2,\dots t_k\}$, as follows:

$$\text{loss}_{stacked} = -\sum_{t=1}^{t_k}\sum_{i=1}^{c} \mathbf{y_t}\log(\hat{\mathbf{y}}_t).$$



# Continue Inference Framework with Buffering

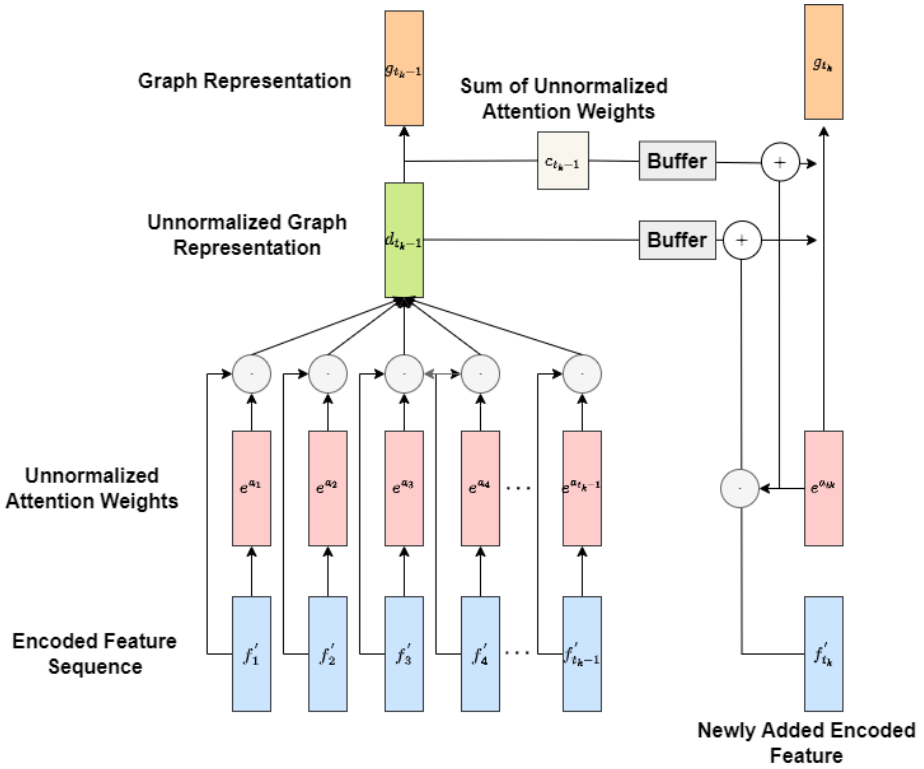
## Continue Inference

- By buffering the running total of the unnormalized weighted features $c_{t_k-1}$ and sum of the unnormalized attention weights $d_{t_k-1}$ at time $t_{k-1}$, we compute the total sum of the feature vectors at the time $t_k$ as follows:

$$\boldsymbol{c_{t_k} = c_{t_{k-1}} + e^{a_{t_k}} f'_{t_k}},$$

$$d_{t_k} = d_{t_k-1} + e^{a_{t_k}}.$$

- In this way, we avoid the redundant calculation of the features and weights from time 1 to $t_k - 1$.



# Experimental Details

## Datasets

- Twitter15 [11]: 374 unverified-rumours, 372 true-rumours, 370 false-rumours, 373 non-rumours
- Twitter16 [11] : 201 unverified-rumours, 207 true-rumours, 205 false-rumours, 205 non-rumours
- CR-Twitter [13]: 143 unverified-rumours, 303 true-rumours, 405 false-rumours, 1334 non-rumours

## Hyperparameters

- We trained the models for 50 epochs with a mini-batch size of 16, which is the largest multiple of 2 that fits the GPU memory in the experiment setup.

- We employ Adam optimizer with an initial learning rate of 2e-5 and a linear learning decay from the initial value to 0.

[11] Ma et al., “Rumour Detection on Twitter with Tree-structured Recursive Neural Networks,” *ACL*,” 2018.
[13] Ke et al., “Novel Approach for Cantonese Rumour Detection based on Deep Neural Network,” *IEEE SMC*, 2020.

# Rumour Verification Results on Twitter15 Dataset

| Model | Accuracy | F1 (Macro) | F1 (NR) | F1 (FR) | F1 (TR) | F1 (UR) |
|---|---|---|---|---|---|---|
| DTC [81] | 0.454 | 0.455 | 0.733 | 0.355 | 0.317 | 0.415 |
| SVM-TS [42] | 0.544 | 0.539 | 0.796 | 0.472 | 0.404 | 0.483 |
| GRU-RNN [28] | 0.641 | 0.644 | 0.684 | 0.634 | 0.688 | 0.571 |
| BU-RvNN [24] | 0.708 | 0.709 | 0.695 | 0.728 | 0.759 | 0.653 |
| TD-RvNN [24] | 0.723 | 0.729 | 0.682 | 0.758 | 0.821 | 0.654 |
| STS-NN [25] | 0.809 | 0.809 | 0.797 | 0.811 | 0.856 | 0.773 |
| PLAN [82] | 0.845 | 0.845 | 0.823 | 0.858 | 0.895 | 0.802 |
| StA-PLAN [82] | 0.852 | 0.852 | 0.840 | 0.846 | 0.884 | 0.837 |
| Bi-GCN [27] | 0.886 | 0.886 | 0.891 | 0.860 | 0.93 | 0.864 |
| PPA-WAE[83] | 0.873 | 0.873 | 0.899 | 0.881 | 0.869 | 0.843 |
| DA-GCN [84] | 0.905 | 0.905 | **0.959** | 0.895 | 0.914 | 0.852 |
| GACL [85] | 0.901 | 0.897 | 0.958 | 0.851 | 0.903 | **0.876** |
| **CDGTN** | **0.916** | **0.915** | 0.947 | **0.951** | **0.912** | 0.859 |

NR: Non-Rumour
FR: False-Rumour
TR: True-Rumour
UR: Unverified-Rumour

# Rumour Verification Results on Twitter16 Dataset

| Model | Accuracy | F1 (Macro) | F1 (NR) | F1 (FR) | F1 (TR) | F1 (UR) |
|---|---|---|---|---|---|---|
| DTC [81] | 0.465 | 0.465 | 0.643 | 0.393 | 0.419 | 0.403 |
| SVM-TS [42] | 0.574 | 0.568 | 0.755 | 0.420 | 0.571 | 0.526 |
| GRU-RNN [28] | 0.633 | 0.609 | 0.617 | 0.715 | 0.577 | 0.527 |
| BU-RvNN [24] | 0.718 | 0.718 | 0.723 | 0.712 | 0.779 | 0.659 |
| TD-RvNN [24] | 0.737 | 0.737 | 0.662 | 0.743 | 0.835 | 0.708 |
| STS-NN [25] | 0.809 | 0.809 | 0.797 | 0.811 | 0.856 | 0.773 |
| PLAN [82] | 0.874 | 0.874 | 0.853 | 0.839 | 0.917 | 0.888 |
| StA-PLAN [82] | 0.868 | 0.869 | 0.826 | 0.833 | 0.927 | 0.888 |
| Bi-GCN [27] | 0.880 | 0.880 | 0.847 | 0.869 | 0.937 | 0.865 |
| PPA-WAE[83] | 0.887 | 0.887 | 0.882 | 0.903 | 0.921 | 0.842 |
| DA-GCN [84] | 0.902 | 0.902 | 0.894 | 0.872 | 0.928 | 0.913 |
| GACL [85] | 0.920 | 0.917 | **0.934** | 0.869 | **0.959** | 0.907 |
| **CDGTN** | **0.929** | **0.927** | 0.874 | **0.949** | 0.956 | **0.931** |

NR: Non-Rumour
FR: False-Rumour
TR: True-Rumour
UR: Unverified-Rumour

# Rumour Verification Results on CR-Twitter Dataset

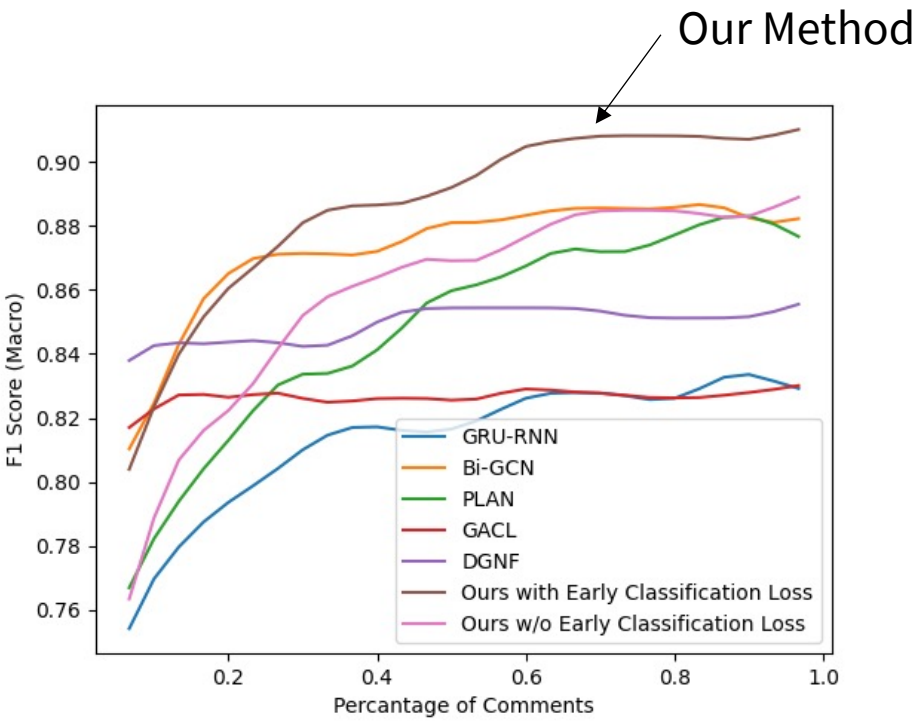
| Model | Accuracy | F1 (Macro) | F1 (NR) | F1 (FR) | F1 (TR) | F1 (UR) |
|---|---|---|---|---|---|---|
| RNN-GRU [28] | 0.841 | 0.778 | 0.903 | 0.744 | 0.726 | 0.739 |
| PLAN [82] | 0.855 | 0.790 | 0.916 | 0.766 | 0.702 | 0.778 |
| GNN-LSTM [86] | 0.855 | 0.721 | 0.928 | 0.765 | 0.790 | 0.400 |
| GAT [87] | 0.871 | 0.825 | 0.922 | 0.772 | 0.815 | 0.793 |
| GCN [27] | 0.841 | 0.757 | 0.911 | 0.770 | 0.667 | 0.682 |
| Bi-GCN [27] | 0.867 | 0.808 | 0.919 | 0.797 | 0.734 | 0.783 |
| GACL [85] | 0.880 | 0.825 | **0.935** | 0.813 | 0.760 | 0.792 |
| DCNF [88] | 0.880 | 0.826 | 0.930 | **0.826** | 0.739 | 0.809 |
| **CDGTN** | **0.891** | **0.858** | 0.931 | 0.824 | **0.805** | **0.873** |

NR: Non-Rumour
FR: False-Rumour
TR: True-Rumour
UR: Unverified-Rumour

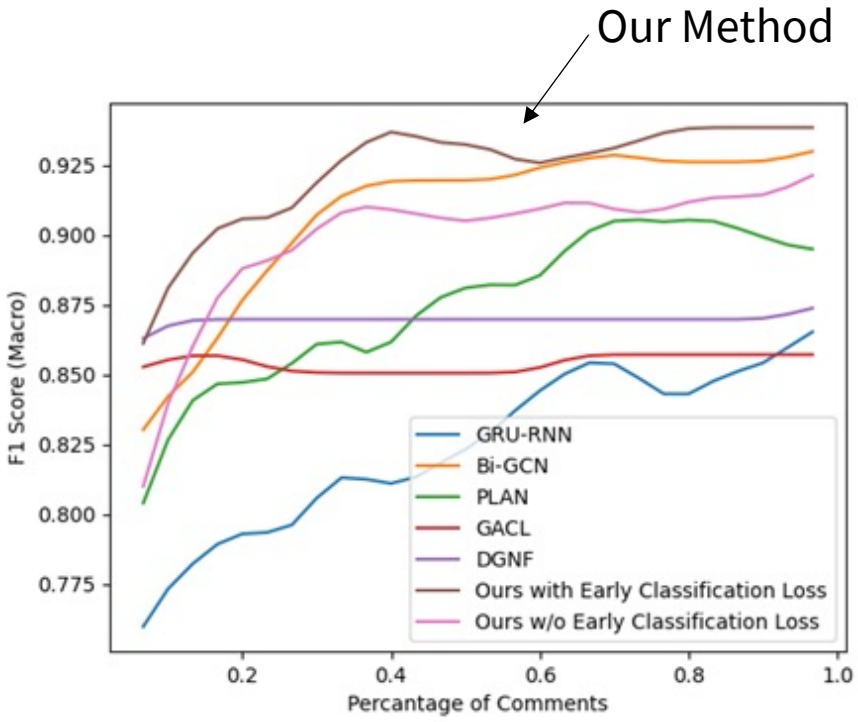
# Ablation Study on CDGTN

| Dataset | Model | Accuracy | F1 (Macro) |
|---|---|---|---|
| Twitter15 | CDGTN w/o attention mask | 0.909 | 0.908 |
| | CDGTN w/o SGIAP | 0.882 | 0.881 |
| | CDGTN (Full) | **0.912** | **0.912** |
| Twitter16 | CDGTN w/o attention mask | 0.933 | 0.932 |
| | CDGTN w/o SGIAP | 0.926 | 0.926 |
| | CDGTN (Full) | **0.939** | **0.939** |
| CR-Twitter | CDGTN w/o attention mask | 0.818 | 0.750 |
| | CDGTN w/o SGIAP | 0.878 | 0.820 |
| | CDGTN (Full) | **0.891** | **0.858** |

# Early Rumour Classification Results on Twitter Datasets

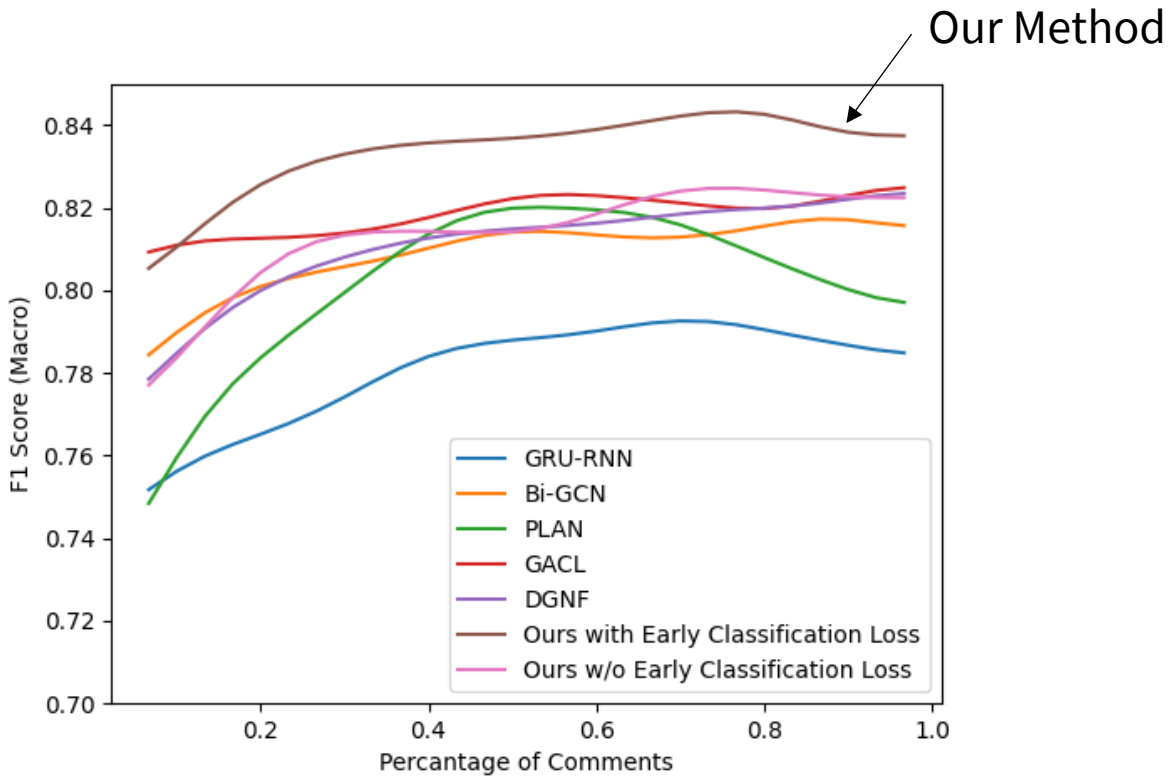



Twitter15 Dataset

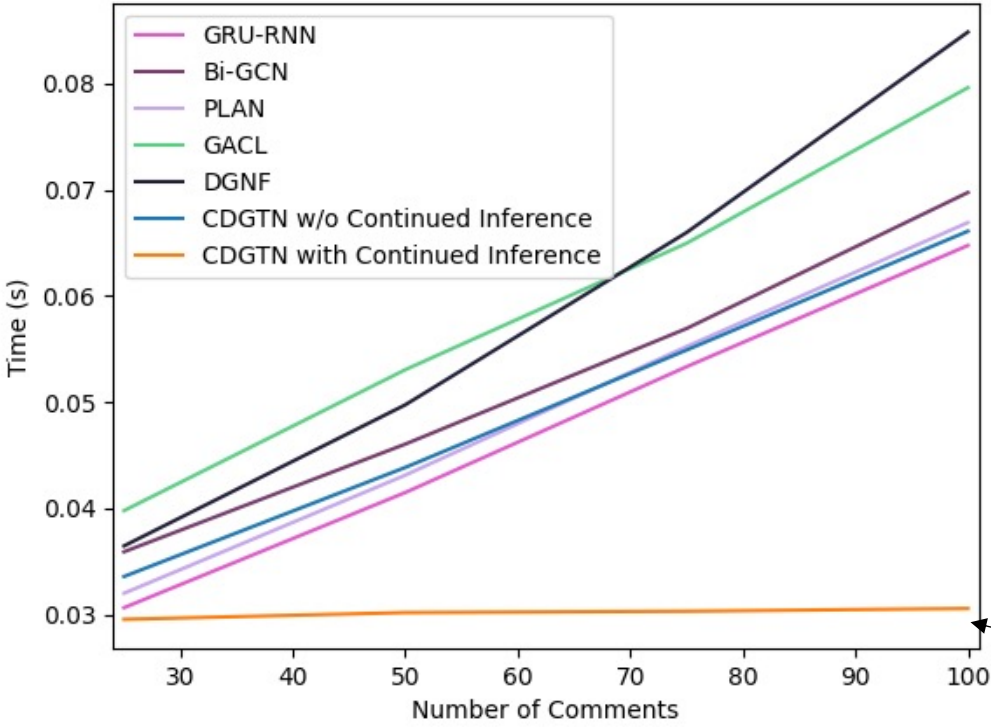



Twitter16 Dataset

# Early Rumour Classification Results on CR-Twitter Dataset



# Time Complexity Analysis of Stream Verification




Our Method

# Examples on CR-Twitter Dataset

| Label | | Message |
|---|---|---|
| False Rumour | Source | **Large-scale protest in Hunan, China** |
| | Replies | Please give a news link, thank you. |
| | | Impossible, absolutely impossible. |
| | | You idiot. See it clearly, don't spread fake news. |
| | | This kind of fake account is not serious. |
| True Rumour | Source | **Playing baseball in a Power failure.** |
| | Replies | This is so funny. |
| | | Does everyone wear night vision goggles? |
| | | The Uni-President Lions still played today, but Su Zhijie lost power halfway through the game. |
| | | I laughed so hard yesterday when watching the live. |
| Unverified Rumour | Source | **At 4 a.m., a fire broke out in a community in Yibin, Sichuan, and fire trucks came, but they couldn't get in, and the iron sheet of closure and control blocked it.** |
| | Replies | No casualties, right? |
| | | True or false? |
| | | Is it true? |
| | | It is going to spread rumours again. |
| Non-Rumour | Source | **On August 22, Zheng Zhongwei, director of the Science and Technology Development Center of the National Health and Medical Commission and head of the Vaccine Research and Development Working Group of the Joint Prevention and Control Mechanism of the State Council, said in the CCTV "Dialogue" program that my country has officially launched the new crown vaccine on July 22. for emergency use.** |
| | Replies | Give the vaccine to African brothers first, they need it more. |
| | | Keep it up! |
| | | In China, epidemic prevention and control and vaccine research and development are so excellent that people feel safe and peaceful. |
| | | Pay tribute to the "rebels" all over the world and may the human disasters on the scene disappear. |

# IV. External Evidence-Based Claim Veracity Assessment using Large Language Models (LLMs)

[6] T. Cheung and K. Lam, “FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking,” in *Proceedings, APSIPA ASC 2023*, Oct. 31, 2023.

## Background

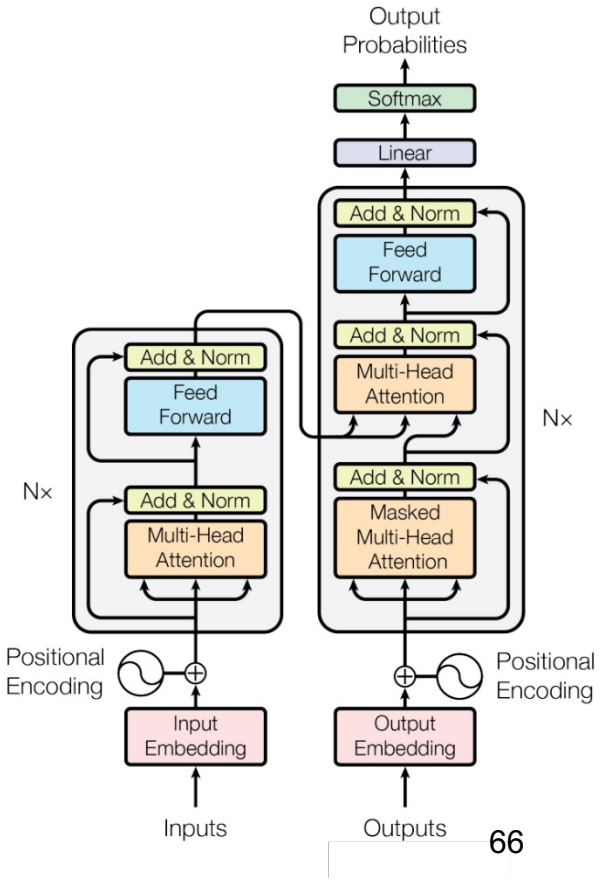
- The emergent Large Language Models (LLMs) has revolutionized the research field of Natural Language Processing (NLP) in 2023.

## Contribution

- We aim to examine the fact-checking capability of open-source LLMs, i.e., LLaMA [14,15], in terms of inherent knowledge, reasoning ability with external evidence, and comparison to the fine-tuned Small Language Models (SLMs) i.e., BERT.

**BERT** Encoder-Only

**LLaMA** Decoder-Only




[14] Touvron et al., "LLaMA: Open and Efficient Foundation Language Models." arXiv, 2023.
[15] Touvron et al., "LLaMA 2: Open Foundation and Fine-Tuned Chat Models." arXiv, 2023.

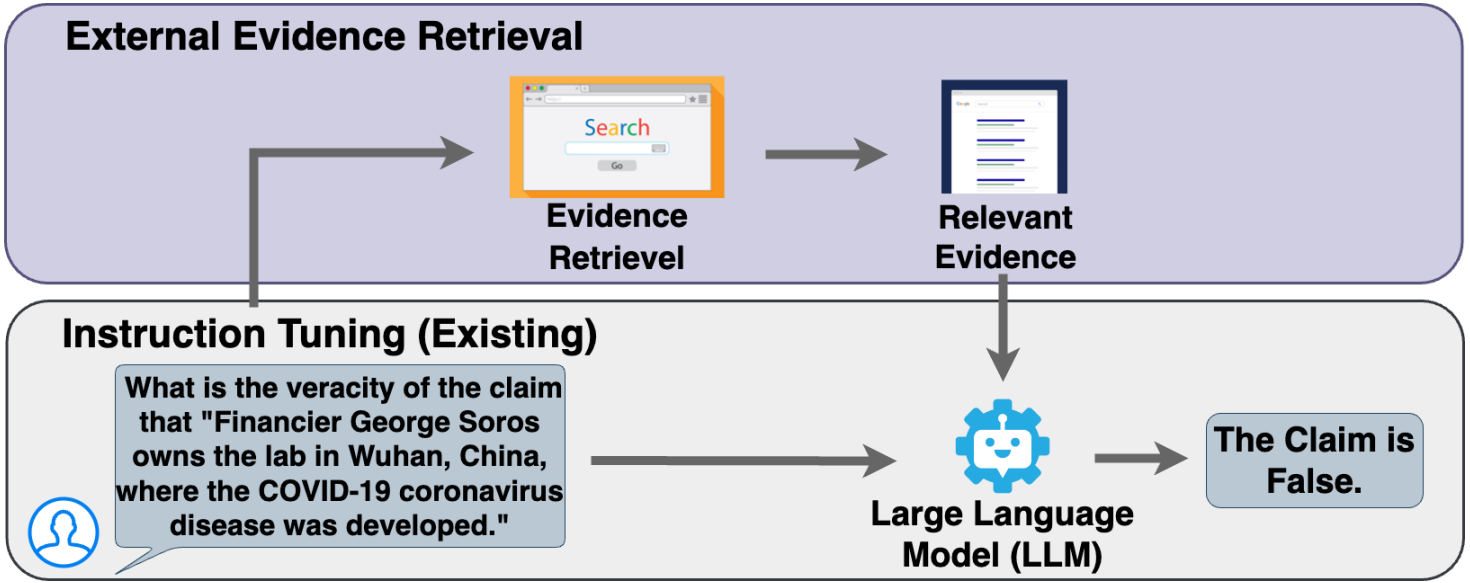
# Research Questions

1. How does the effectiveness of fact-checking using LLMs compare between inherent knowledge and external knowledge utilization?
2. Does fine-tuning LLMs lead to superior fact-checking performance, and how does it compare between models with and without external knowledge?
3. In the context of fact-checking, how does the performance of LLaMA-based fine-tuning compare to BERT-based fine-tuning?

LLAMA 2

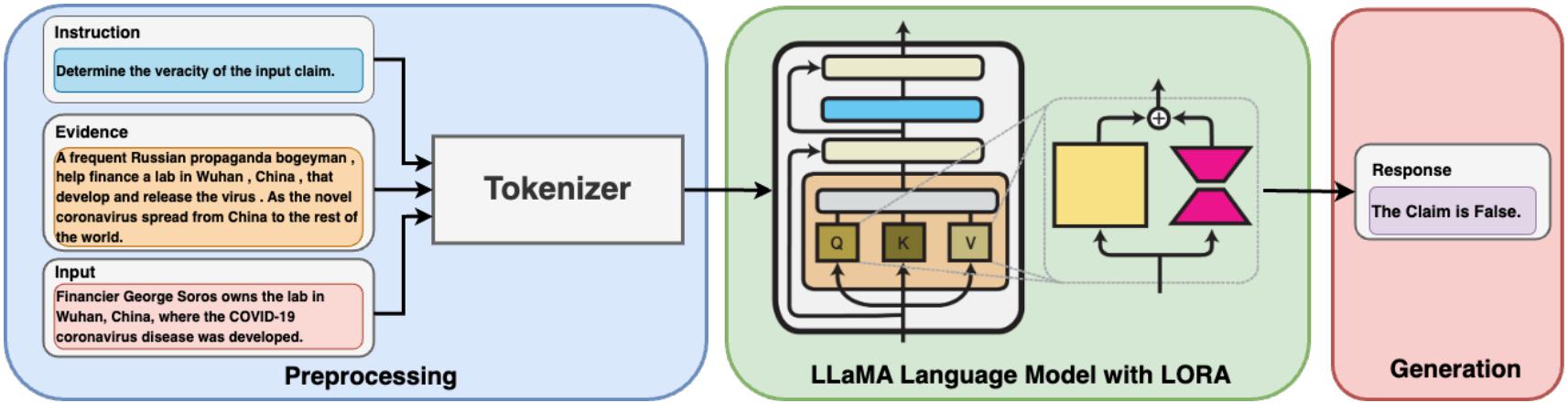
# External Evidence-Based Claim Veracity Assessment

- To improve the fact-checking capability, we adopt the Google Search API to retrieve relevant information to verify a statement.



# Instruction-Tuning with External Evidence

- Due to the limited computational power, we adopt Low Rank Adaption (LORA) [15] and 8-bit quantization [17] to fine-tune LLaMA with external evidence for automatic fact-checking.




[16] Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models." *Int. Conf. on Learning Representations*. 2021.
[17] Dettmers et al. "Llm.int8(): 8-bit matrix multiplication for transformers at scale." *NeurIPS, 2022.*

# Experimental Details

## Implementation Details

- LLaMA-7B [14] is used as the backbone model for fine-tuning.
- LLaMA2-Chat-7B [15] is used for zero-shot prompting.

| Hyperparameter | Value |
|---|---|
| Number of Epoch | 3 |
| Mini-batch Size | 2 (with gradient accumulation 16x) |
| Optimizer | Adam Optimizer |
| Initial Learning rate | Linearly decay to 0 |



[14] Touvron et al., “LLaMA: Open and Efficient Foundation Language Models.” arXiv, 2023.
[15] Touvron et al., “LLaMA 2: Open Foundation and Fine-Tuned Chat Models.” arXiv, 2023.

# Comparison to Other SOTA Methods

| Methods | Precision | Recall | F1-Score |
|---|---|---|---|
| SVM [104] | 0.3233 | 0.3251 | 0.3171 |
| CNN [101] | 0.3880 | 0.3850 | 0.3859 |
| RNN [105] | 0.4135 | 0.4209 | 0.4039 |
| DeClarE [106] | 0.4339 | 0.4352 | 0.4218 |
| dEFEND [97] | 0.4493 | 0.4326 | 0.4407 |
| sentHAN [107] | 0.4566 | 0.4554 | 0.4425 |
| SBERT-FC [34] | 0.5106 | 0.4592 | 0.4551 |
| GenFE [99] | 0.4429 | 0.4474 | 0.4443 |
| GenFE-MT [99] | 0.4564 | 0.4527 | 0.4508 |
| CofCED [98] | 0.5299 | 0.5099 | 0.5107 |
| **FactLLaMA** | **0.5611** | **0.5550** | **0.5565** |

# Fine-tuned vs Zero-Shot Prompt-Based Methods

| Settings | Methods | Precision | Recall | F1 |
|---|---|---|---|---|
| BERT-Based Fine-Tuning | SBERT-FC | 0.5106 | 0.4592 | 0.4551 |
| | GenFE | 0.4429 | 0.4474 | 0.4443 |
| | GenFE-MT | 0.4564 | 0.4527 | 0.4508 |
| | CofCED | 0.5299 | 0.5099 | 0.5107 |
| LLaMA-Based Zero-Shot Prompting | LLaMA2-Chat* (w/o external knowledge) | 0.4198 | 0.4311 | 0.3775 |
| | LLaMA2-Chat* (with external knowledge) | 0.4881 | 0.4858 | 0.4793 |
| LLaMA-Based Fine-Tuning | FactLLaMA (w/o external knowledge) | 0.5376 | 05400 | 0.5376 |
| | **FactLLaMA (with external knowledge)** | **0.5611** | **0.5550** | **0.5565** |

*Experiments with LLaMA2-Chat were done after submitting the thesis.

# Examples of External Evidence-Based Fact-Checking

| Label | Text | |
|---|---|---|
| False | Source | An ISIS flag is being displayed in the window of a café under siege in Sydney's Martin Place. |
| | External Evidence | The flag displayed during the siege at a Sydney cafe is not the same one used by the Islamic State terrorist group. |
| | | The black flag with white writing hung in the window was initially mistaken by many for an Isis flag. |
| | | Siege makes global headlines. The flag shown being held by hostages against the window of Lindt Chocolat Cafe is not an Islamic State flag but an Islamic flag that has been co-opted. |
| True | Source | Germanwings Airbus A320 crashes in French Alps |
| | External Evidence | Germanwings A320 aircraft flying from Barcelona to Düsseldorf goes down in southern French Alps with 150 on board ... German Airbus A320 plane crashes in French Alps. |
| | | A Germanwings plane carrying 150 people has crashed in the French Alps on its way from Barcelona to Duesseldorf. The Airbus A320 - flight 4U 9525 - went down between Digne and Barcelonnette. |
| | | An Airbus A320 with 144 passengers and 6 crew members has crashed in Digne region. |

# Summary

## RQ1: External Knowledge Boost

- Incorporating external knowledge improves LLM-based fact-checking performance. LLaMA2-Chat with external knowledge enhances overall effectiveness.

## RQ2: Fine-Tuning Success

- Fine-tuning LLMs, especially with external knowledge, boosts fact-checking performance. Fine-tuned FactLLaMA demonstrates heightened effectiveness.

## RQ3: Fine-tuned LLaMA beats Fine-tuned BERT

- LLaMA-based fine-tuning excels in fact-checking over BERT, showcasing superior performance. FactLLaMA outperforms SBERT-FC in overall effectiveness.

# Conclusion and Future Work

# Comparison of Different Methods

| Method | Advantages | Disadvantages |
|---|---|---|
| **Multimodal Image-Text Rumour Detection** | Comprehensive understanding with text and visual integration. | Reliable labeled dataset needed for training. |
| | Deep learning with crossmodal attention macahsim captures complex patterns. | Dependency on image availability and quality. |
| **Simultaneous Rumour and Malicious Account Detection** | Early detection approach to find misleading information. | Identification may require additional metadata. |
| | User Profiling in identifying accounts. | Overreliance on behavior patterns may lead to false results. |
| **Propagation Graph-Based Claim Veracity on Social Media** | Graph-based approach captures spread and propagation patterns. | Constructing the holistic propagation graphs is computationally expensive. |
| | Incorporation of community enhances veracity assessment. | Rely on availability and latency of network and community response. |
| **Claim Veracity Assessment using External Evidence** | External evidence provided for broader perspective. | Reliability of external sources and fact-checking reports may vary. |
| | Access to fact-checking reports and expert opinions improves reliability. | Challenge of effectively integrating diverse evidence. |

# Conclusion

## Holistic Investigation for Rumour Detection and Veracity Assessment

- Innovative approaches, including multimodal image-text classification, author-aware rumour detection, propagation graph-based analysis, and external evidence-based claim veracity assessment.

## Addressing Challenges in Misinformation

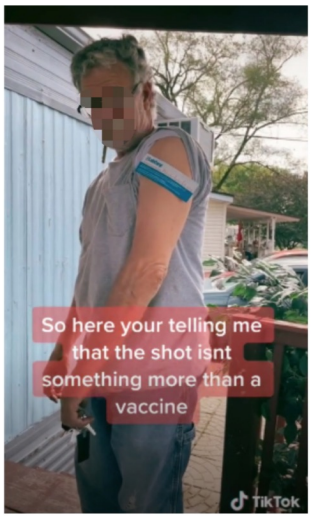
- Addressing challenges posed by rumours and false information, showcasing the potential for positive impacts on individuals, society, and the overall information ecosystem.

## Call for Future Research and Ethical Considerations

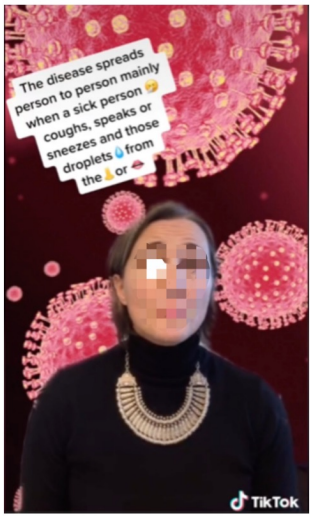
- Encourage future research, such as LLMs, to build upon these findings, considering emerging trends and ethical considerations.

# Future Work

- **Multimodal Fake Video Detection**. Extend the proposed methods to video-sharing platforms, such as YouTube [18], TikTok [19], and DouYin [20], using audio, video, voice-to-text features.




(a) Misleading



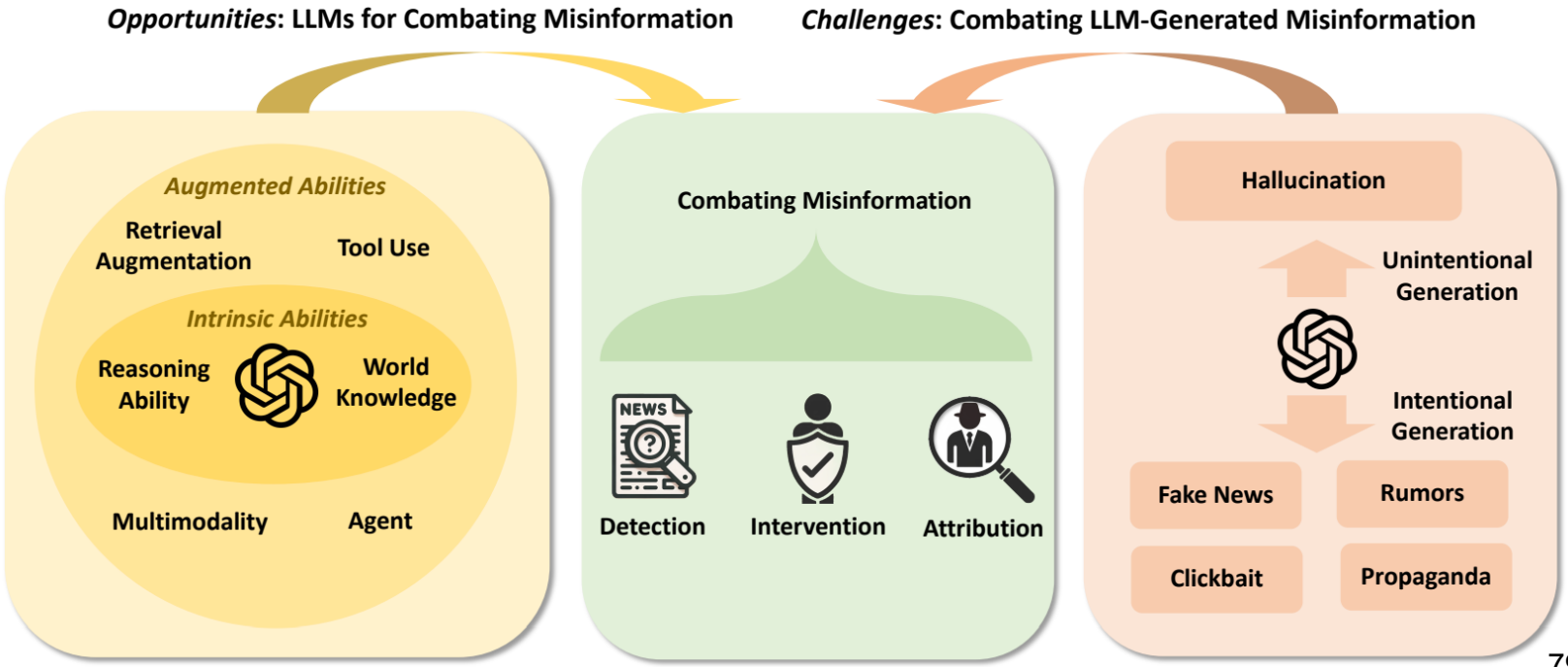

(b) Non-misleading

**Video (a)** is a misleading video that shows a man whose arm can attract a magnetic strip after receiving a shot of COVID-19 vaccine. **Video (b)** is a non-misleading video that presents how COVID-19 spreads in our daily life. (Faces are blurred in the images for user privacy concern.)

[18] H. Choi and Y. Ko, “Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection,” *CIKM*, 2021.
[19] Shang et al., "A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok," *IEEE Int. Conf.on Big Data*2021, pp. 899-908 .
[20] Qi et al., “FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms,”, *AAAI*, vol. 37, no. 12, pp. 14444–14452, Jun. 26, 2023.

# Future Work

- **Hallucination Detection.** Address the problem of LLM-Generated Misinformation.






[21] C. Chen and K. Shu, "Combating Misinformation in the Age of LLMs: Opportunities and Challenges." *arXiv*, 2023.

# References


[1] Rohera *et al.*, “A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects,” *IEEE Access*, vol. 10, 2022.

[2] Zubiaga et al., “Detection and Resolution of Rumours in Social Media,” *ACM Computing Surveys*, vol. 51, no. 2, pp. 1–36, 2018.

[3] T. Cheung and K. Lam, “Crossmodal bipolar attention for multimodal classification on social media,” *Neurocomputing*, vol. 514, pp. 1–12, Dec. 2022.

[4] T. Cheung and K. Lam, “Author-Aware Rumour Detection with Layer-Wise Parameter-Efficient Tuning and Incomplete Feature Learning,” submitted to *IEEE Access.*

[5] T. Cheung and K. Lam, “Causal diffused graph-transformer network with stacked early classification loss for efficient stream classification of rumours*,” Knowledge-Based Systems*, vol. 277, pp. 110807, 2023.

[6] T. Cheung and K. Lam, “FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking,” in *Proceedings, APSIPA ASC 2023*, Oct. 31, 2023.

[7] Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *ICLR*, 2021.

[8] Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, 2019.

[9] Augbiaga et al., “Exploiting Context for Rumour Detection in Social Media*,” International Conference on Social Informatics*,” 2017.

[10] Jin et al., “Multimodal Fusion with Recurrent Neural Networks for Rumour Detection on Microblogs,” *ACM MM*, 2017.

# References


[11] Ma et al., “Rumour Detection on Twitter with Tree-structured Recursive Neural Networks,” *ACL*,” 2018.

[12] Chen et al., “Identifying Cantonese rumors with discriminative feature integration in online social networks,” *Expert Systems with Applications*, 2020.

[13] Ke et al., “Novel Approach for Cantonese Rumour Detection based on Deep Neural Network,” *IEEE SMC*, 2020.

[14] Touvron et al., “LLaMA: Open and Efficient Foundation Language Models.” arXiv, 2023.

[15] Touvron et al., “LLaMA 2: Open Foundation and Fine-Tuned Chat Models.” arXiv, 2023.

[16] Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations*. 2021.

[17] Dettmers et al. "Llm. int8 (): 8-bit matrix multiplication for transformers at scale." *NeurIPS, 2022.*

[18] H. Choi and Y. Ko, “Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection,” *CIKM*, 2021.

[19] Shang et al., "A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok*," IEEE International Conference on Big Data (Big Data)*, 2021, pp. 899-908 .

[20] Qi et al., “FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms,”, *AAAI*, vol. 37, no. 12, pp. 14444–14452, Jun. 26, 2023.

[21] C. Chen and K. Shu, “Combating Misinformation in the Age of LLMs: Opportunities and Challenges.” *arXiv*, 2023.

# Thank You